RICE UNIVERSITY

Genomic Detection Using Sparsity-inspired Tools

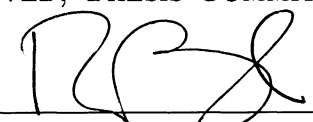
by

Mona A. Sheikh

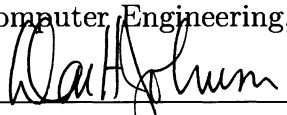
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

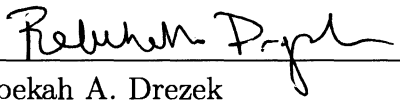
APPROVED, THESIS COMMITTEE:



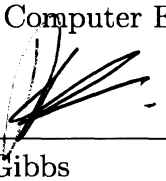
Richard G. Baraniuk, Chair
Victor E. Cameron Professor of Electrical & Computer Engineering, Rice University



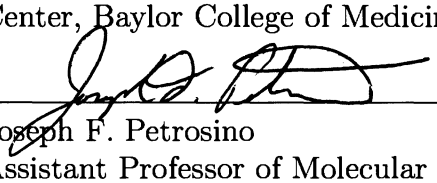
Don H. Johnson
J.S. Abercrombie Professor Emeritus of Electrical & Computer Engineering, Rice University



Rebekah A. Drezek
Professor of Bioengineering, and Electrical & Computer Engineering, Rice University



Richard A. Gibbs
Wofford Cain Professor of Molecular & Human Genetics
Director, Human Genome Sequencing Center, Baylor College of Medicine



Joseph F. Petrosino
Assistant Professor of Molecular Virology & Microbiology
Baylor College of Medicine

Houston, Texas
November, 2010

ABSTRACT

Genomic Detection Using Sparsity-inspired Tools

by

Mona A. Sheikh

Genome-based detection methods provide the most conclusive means for establishing the presence of microbial species. A prime example of their use is in the detection of bacterial species, many of which are naturally vital or dangerous to human health, or can be genetically engineered to be so. However, current genomic detection methods are cost-prohibitive and inevitably use unique sensors that are specific to each species to be detected. In this thesis we advocate the use of combinatorial and non-specific identifiers for detection, made possible by exploiting the sparsity inherent in the species detection problem in a clinical or environmental sample. By modifying the sensor design process, we have developed new molecular biology tools with advantages that were not possible in their previous incarnations. Chief among these advantages are a universal species detection platform, the ability to discover unknown species, and the elimination of PCR, an expensive and laborious amplification step prerequisite in every molecular biology detection technique. Finally, we introduce a sparsity-based model for analyzing the millions of raw sequencing reads generated during whole genome sequencing for species detection, and achieve significant reductions in computational speed and high accuracy.

Acknowledgments

I would like to thank my advisor, Richard Baraniuk for his guidance and encouragement, especially while I ventured into unknown multidisciplinary territory, and for teaching me to think "big". I also thank my committee members: Rebekah Drezek, Richard Gibbs, Don Johnson and Joseph Petrosino for all their support and time. Thanks also go to my experimental collaborators: all wet lab experiments at Rice were conducted with the help of Adam Lin, bioengineering graduate student in the Drezek lab at Rice University, and DNA extractions from bacteria were performed by Bonnie Youmans, molecular biology/virology graduate student at Baylor College of Medicine. Finally, significant thanks are extended to my family – my parents continually inspire my interest in the health sciences and innovation – and friends who helped make this possible.

Contents

Abstract ii

List of Illustrations vii

List of Tables 1

1 Introduction 2

1.1 DNA-based Bacterial Identification 5

2 Background 8

2.1 Compressed Sensing 8

2.1.1 Sparse Approximation 10

2.2 Genomic tools 10

2.2.1 DNA Microarrays 12

2.2.2 Molecular Beacons 14

2.2.3 Sequencing 15

2.3 DNA Hybridization Models 17

2.3.1 DNA Hybridization Affinities via Smith-Waterman Alignment 18

2.3.2 DNA Hybridization Affinities via a Nucleic Acid Thermodynamics Model 19

3 Compressed Sensing Microarrays 22

3.1 Concerns with Traditional DNA Microarrays 22

3.2 Designing Compressed Sensing Microarrays 23

3.3 Target Assignment and Probe Selection Algorithm 25

3.4 Using Smith-Waterman Alignment to predict Φ 27

3.5 Decoding Nonlinear Measurements via Belief Propagation 28
3.6 Problems with COGs-based Probe Design 31

4 Nonspecific Sensing via Random Probes 33
4.1 Need for nonspecificity in bacterial detection 33
4.1.1 Genome-based bacterial detection 35
4.1.2 Random probes powered by Compressed Sensing 37
4.2 Random Probe Microarrays 39
4.3 Implications of a Nonspecific RPM 42
4.4 Simulations in silico: Bacterial Detection by Random Probes 45
4.4.1 Hybridization affinity generation for Φ 46
4.4.2 Trends in bacterial detection using CS reconstruction algorithms 47
4.5 Experimental Design 53
4.5.1 Experimental calibration of Φ 54
4.5.2 Random Molecular Beacon Experiments 55
4.5.3 Extensions to other molecular devices 61
4.6 Real world considerations 62
4.6.1 Application scenarios 62
4.6.2 Perturbations to Φ 65

5 Sparsity-based Methods for Bacterial Sequencing Data 69
5.1 Motivation 69
5.2 Proposed solution 70
5.3 Previous work 72
5.4 Linear Sparse Approximation Setup 73
5.5 Kmer Frequencies in Genomes 75
5.6 Using Kmer-based Φ's for Detection 83
5.7 Implications of Sparsity-based Kmer Analysis in Sequencing 86

6 Conclusions **89**

Bibliography **94**

Illustrations

3.1 Structure of sensing matrix Φ of a CSM with M spots identifying N targets 24

3.2 Plot of normalized L_2 measurement error vs. number of measurements comparing decoding nonlinear measurements by our modified BP version and BP that ignores the nonlinearity. Number of signal coeffs $N = 200$; $\alpha = \beta = 25$; $\sigma_y = 2$ 30

4.1 ROC curve showing Probability(detection) vs. Probability(false positive) for random probes of length 19. The two curves in yellow and black are for the species Onions yellow and Ehrlichia ruminantium, which have small genomes and low GC content. Onions yellow performs even worse than E. ruminantium due to its smaller genome. 49

4.2 ROC curve showing Probability(detection) vs. Probability(false positive) for random probes of length 21. The two curves in yellow and black are for the species Onions yellow and Ehrlichia ruminantium, which have small genomes and low GC content; their detection here is worse than for length 19. 50

4.3 ROC curve showing Probability(detection) vs. Probability(false positive) for random probes of length 23. The two curves in yellow and black are for the species Onions yellow and Ehrlichia ruminantium, which have small genomes and low GC content; their detection is slightly improved but shows a high false positive rate due to sloppy binding. 51

4.4 Locations of hybridization for a sample random probe of length 19 in an E. coli genome. Percentage values were omitted in places for clarity, but the bands specify their locations. 54

4.5 General depiction of combinatorial probe-based sensing in the RPM, with affinities in percentages in Φ from a thermodynamic hybridization model. Each percentage refers to the affinity of a fixed molarity of a target with a fixed molarity of probe. 56

4.6 Depiction of the experimental calibration of values of Φ from fluorescent intensities in RFU's (relative fluorescence units). There may be a slight deviation from the intensity predicted by theory compared with that in an experiment due to the negative control being added multiple times for each element of Φ. 57

4.7 Probability of species detection varying with the number of probes (M) using a thresholded at 20%, binary Φ as our perturbed Φ for decoding. This plot shows the detection curves for 20 randomly chosen bacteria out of $N = 100$. Notice that the only species that cannot be detected at $M = 200$ in this set of 20 is E. ruminantium with GC content of 27%. 63

4.8 Here Φ values are thresholded at 50% and then converted to binary. We see decreased detection performance due to greater perturbation; now only 8 of the 20 bacteria can be detected with the same set of probes. ϵ_Φ here is 0.82, compared to 0.40 in Figure 4.7. 65

5.1 Variation of the KL distance between the kmer frequency estimates of the genomes of E. coli and Aster Yellow Witches' Broom (AYWB) from their random iid counterparts. AYWB has a much smaller genome generating fewer samples than E. coli for a given kmer length, and diverges faster and greater from its iid distribution. . . . 75

5.2 (a) Grammian of the Φ based on 10mer estimates for a set of 40 randomly chosen bacteria (b) Grammian of the Φ based on 10mer estimates with its iid 10mer estimates subtracted. 76

5.3 Grammian of the Φ based on 25mer estimates for the same set of 40 bacteria. 77

5.4 (a) Grammian of the first 10,000 rows with lowest T content of the 25mer Φ (b) Grammian of the last 10,000 rows with highest T content of the 25mer Φ 78

5.5 Detection probability increases with kmer length, for changing sequencing error rates of 5, 10, 20, 30 and 40% 80

5.6 Detection probability increases with kmer length, for changing error magnitude variances of 3, 9, 15, 21 and 27. Variance of 3 corresponds to an SNR of 2dB. 81

5.7 Increasing storage size in MB of Φ with kmer length. 84

Tables

4.1 GC contents and genome lengths of bacteria with lowest detection rates 53

Chapter 1

Introduction

Genomic data is ripe for the plucking by signal processors. With the rapid advance of sequencing technologies, we are quickly acquiring complete, exact representations of the genomes of many organisms. In signal processing, the exact representation of signals by their samples and the Nyquist-Whitaker-Shannon (N-W-S) sampling theorem that first prescribed the theory for it were the starting point for a multitude of signal processing tasks, algorithms and devices. The complete sequencing of genomes offers the scope to apply such intelligence from the last 60 years of signal processing toward the processing of genomic data. The appropriation of signal processing analysis in genomics and genetics will have impact in a variety of fields vital to human life: personalized medicine, drug development, defense, environmental monitoring and food safety, among others.

While signal processing methods can be used to better process the data provided to us by existing technologies, it can also have a reverse influence: to engineer better instruments in the first place. If data analysis can inform data acquisition, it maximizes the overall efficiency of the process. This is not a typical chain of command, since it is invariably more cost-effective to develop new computational methods to analyze some given data than to rebuild the data source itself.

However, recent advances in signal processing theory have created advantages that bolster this reverse influence for practical implementation. The theory of Compressed Sensing (CS) leverages the *sparsity* of signals, and requires far fewer data to be

acquired while performing the same signal processing task with additional gains, including greater efficiency and robustness to error. First described in 2005, CS theory is essentially an important special case of the N-W-S sampling theorem. If a signal (or some representation of it) is sufficiently sparse, it (or its corresponding sparse representation) requires far fewer samples to be exactly reconstructed than the original sampling theorem would indicate. As it turns out, most natural signals have sparse representations; this prescribes a literal universe of signals where CS principles may be applied.

In this thesis we investigate the application of sparsity in the detection of bacterial genomes. We discuss its application on both the sensing and analysis sides of the bacterial detection process. Our innovations, analysis and recommendations are made in the context of existing experimental molecular biology tools and techniques, and attempt to stay true to that reality. While we choose specific device frameworks to describe our sensing concepts, they are not limited to either a particular sensing device, specific molecule to be sensed, or specific organism.

On the sensing side, we argue for a shift away from specificity in bacterial sensor design, towards the design of two new types of sensors: the first is *combinatorial* sensors, which sets the stage for our main focus, *nonspecific* sensors. Both these sensor designs hinge on the mathematics of Compressed Sensing, which tells us that it is possible to recover and recognize a target signal by taking just a few holistic measurements of it, provided it satisfies the notion of sparsity. For our purposes, we recast bacterial identification as a sparse signal reconstruction problem, providing the justification for a CS-based framework, and the design of sensors that take measurements as CS theory dictates.

Combinatorial sensors are specifically designed to detect a group of specific targets;

we lay out a design for them in the context of microarray probes in Chapter 3. They are able to minimize the number of sensors needed by leveraging CS principles. Nonspecific sensors, on the other hand, exploit another idea from CS theory – beyond minimizing the number of sensors required by also sensing combinatorially, they allow for a *random* flavor in their design and are not created for any particular target or group of targets. We describe the design of nonspecific sensors in the context of microarray probes and molecular beacons in Chapter 4.

Nonspecific sensor design proposes a paradigm shift in sensing ideology: we no longer need know what we are looking for a priori to the sensing procedure. This is in contrast to traditional designs, where – analogous to the man looking for his wallet under the streetlamp, irrespective of where he actually lost it – we are constrained to "see" by what our sensors specify.

On the analysis side, we describe a sparsity-based model for the fast processing of Whole Genome Sequencing data from bacteria. This is a difficult problem where we have a surplus of sequencing data that we would like to analyze quickly to identify bacteria present, without sacrificing (perhaps even improving) accuracy. Current analysis times exceed actual data acquisition times, so there is an immediate need for signal processing innovations to improve the state-of-the-art, particularly for multiple organism detection.

The thesis is organized as follows. In the rest of Chapter 1 we set the stage for genome-based bacterial detection. In Chapter 2 we describe some background material in molecular biology tools, DNA hybridization models and CS. In Chapter 3 we introduce the idea of combinatorial sensors in Compressed Sensing Microarrays, where each sensor detects a group of targets. Here we also outline steps to design a CSM, and a CS decoding algorithm that accounts for nonlinearities and noise in the

measurements. The idea of combinatorial sensing segues into Chapter 4, where we first introduce the idea of nonspecific sensing via random probes. We propose two forms for it: Random Probe Compressed Sensing Microarrays, and Random Molecular Beacons; for the latter we discuss some preliminary experimental work. In Chapter 5 we discuss the identification of bacteria from sequencing data. Finally, in Chapter 6 we conclude with the summarized implications of our work.

1.1 DNA-based Bacterial Identification

Bacterial identification is important in many critical scenarios, from health centers to food processing plants to battlefields. However, the current methodology used to accomplish this in such practical situations is highly primitive. In clinical setups for instance, the primary method for identification even today remains the culturing and then visual inspection of a sample – slow and painstaking steps. In other fields, DNA-based tools often take the backseat to other protein-based tools that look for the secondary compounds that bacteria may secrete in order to identify them.

The chief roadblocks to the widespread practical adoption of DNA-based detection tools stem from their extended processing time and heavy cost in terms of tools and reagents. All the above methods rely on "sensors" that identify the *unique* identifiers in bacterial genomes. The bacterial identifiers that are used reside in the 16S and 23S ribosomal DNA genes in bacteria – stretches of just 2000-3000 base pairs in a genome that is typically on the order of several million [1]. These sequences show significant variation between bacterial species, and serve as good indicators for organism presence. However, several factors in their detection procedure also render them the Achilles' heel of current DNA-based bacterial identification methods.

There are three main problems that result from the use of these short, unique

identifiers. One, due to how short the 16S and 23S genes are, every sample must undergo a lengthy PCR step to amplify the 16S or 23S regions in order for the unique sensors (or probes, used interchangeably in this thesis) to identify them. Two, even after PCR, current molecular tools do not have the capacity to hold all the unique identifiers needed to detect all the different sequenced bacterial species, which means that several different devices each tailored to a different set of bacteria may be needed for a complete detection. Three, current molecular tools have no way to identify unknown or mutated species with new 16S or 23S sequences that are not contained in their set of identifiers. Studies of prokaryotic diversity estimate that there are yet millions of undetected bacterial species in the world, of which fewer than 3000 have completely sequenced genomes at last count. By focusing solely on the 16S or 23S known identifiers in each bacteria, current molecular tools are severely limited in their detection capabilities.

The nonspecific probe design methods based on Compressed Sensing that we describe promise relief in alleviating the above disadvantages. One, the probes in the nonspecific design capture holistic group-based DNA measurements across the entire bacterial genome instead of focusing on a single 16S or 23S region. This paves the path to detection methods that do not need PCR. Two, since each sensor can measure a group of targets instead of a single one, it implies that far fewer sensors are needed to create a mapping for the entire set of bacteria. Therefore there is a reduction in genomic measurements, while maintaining accuracy and efficiency. The group-based detection method also confers the advantage of robustness to error, since no single sensor-type is solely responsible for a bacterium's detection. Three, due to the random design element of nonspecific probes, they have a universal quality; meaning that the same probes can record a pattern due to any bacteria – including those that

are not yet known – making it "future-proof". This pattern can be used to point to the bacteria that are the closest relatives of the unknown bactreium, and once that bacteria is actually sequenced and added to the model, can be identified.

Combinatorial and nonspecific sensor design ideas and the corresponding data analysis using Compressed Sensing algorithms readily apply to DNA hybridization-based molecular devices available today such as microarrays, molecular beacons or PCR analysis. Fewer measurements translate to fewer sensors; so if each sensor's effective cost is high, as it is for genomic tools, lower costs may also help in greater feasibility of that technology's adoption in different domains. Otherwise, bench-top prices do not readily lend such tools to bedside utility.

But sparsity may show its true power in the domain of sequencing technologies, particularly Next Generation Sequencing (background described in Chapter 2), which contributes terabytes of data in measurements and needs extensive processing times. NGS technologies are still in their infancy and are very expensive. But as their cost drops in the next couple years to be comparable to other genomic tools, solutions to the practical problems that plague it will gain importance. Exploiting the sparsity in the problem to be solved offers a solution to the data deluge synonymous with sequencing, and potentially faster analysis methods. In Chapter 6, our sparsity-based model for bacterial detection using sequencing data is presented with these issues in mind.

Chapter 2

Background

2.1 Compressed Sensing

Compressed Sensing (CS) is a recently developed sampling theory for sparse signals. Its core intuition is that if a signal has only a few nonzero (and many zero) values, it should require correspondingly fewer measurements to reconstruct it.

The main result of CS, introduced by Candes, Romberg, and Tao [2] and Donoho [3], is that a length-N signal x that is K-sparse in some basis can be recovered *exactly* from just $M = O(K \log(N/K))$ measurements of the signal. If we choose the canonical basis, x has $K \ll N$ nonzero and $N - K$ zero entries.

In matrix notation, we obtain a linear set of measurements,

$$y = \Phi x, \tag{2.1}$$

where x is the $N \times 1$ sparse signal vector we aim to sense, y is an $M \times 1$ measurement vector, and the *measurement matrix* Φ is an $M \times N$ matrix. In the presence of measurement noise, the model becomes $y = \Phi x + n$ where n is the noise vector. Since $M < N$, the system is underdetermined and recovery of the signal x from the measurements y is ill-posed.

CS theory has shown that exact recovery of the solution to this problem is possible under two critical conditions for Φ and x: (i) the vector x to be sensed is sufficiently sparse and (ii) the rows of Φ are sufficiently incoherent with the signal sparsity basis. Incoherence is achieved if Φ satisfies the so-called Restricted Isometry Property

(RIP) [4]. For example, random matrices built by sampling Gaussian and Bernoulli distributions satisfy the RIP with high probability.

Under these two conditions, the solution we seek can be shown to be the solution to the ℓ_1 minimization problem,

$$\min||x||_{l_1} \text{ such that } y = \Phi x$$

which is a convex problem that can be solved via a linear program. In fact, a variety of reconstruction methods have been developed to recover sparse x from the measurements y. Besides convex relation methods, such as Basis Pursuit Denoising [5] and Iterative Reweighted ℓ_1 Minimization [6], which solve optimization problems, many greedy methods such as Orthogonal Matching Pursuit [7] and CoSaMP [8] are popular. When Φ itself is sparse, Belief Propagation and related graphical algorithms can also be applied for fast signal reconstruction [9]. There also exist algorithms specifically for situations where the signal x has *structured sparsity* [10], or where sparsity can be assumed in the error model [4].

The benefits of CS are not just in deriving an exact solution of an underdetermined system in specific situations, but extend to the "sensing" side as well. By using an essentially "random" measurement system with incoherence properties to do so, our sensing system is fairly robust against error, and our measurements have a democratic quality to them – meaning that all measurements have equal say in the recovery process, and the loss of any one measurement does not handicap the system anymore than the loss of another.

An important property of CS is its *information scalability*; CS measurements can be used for a wide range of statistical inference tasks besides signal reconstruction, including estimation, detection and classification. In this thesis we consider the problem of signal detection and estimation, but by posing it as a sparse reconstruction

problem.

2.1.1 Sparse Approximation

When the system is overdetermined instead of underdetermined, but the signal x to be recovered is still sparse, the problem is simply described as a sparse approximation problem. The measurement matrix in this case often does not satisfy any particular properties, and in equation 2.1, $M > N$. Traditionally the solution to this problem is the least-squares solution. However, under the assumption of sparsity, other methods such as LASSO or sparsity-based recovery algorithms for Compressed Sensing show further improved results.

2.2 Genomic tools

The genome of an organism refers to its collective genetic material consisting of DNA molecules – literally "acid" molecules in the nucleus of every cell. Incredibly, DNA naturally lends itself to a discrete structural description of its subunits, with respect to an ordering that is of great biological consequence. Each molecule of DNA is comprised of a linear sequence of discrete *nucleotides* – a sequence of nucleic bases against a sugar-phosphate backbone. (Base and nucleotide may be used interchangeably in this thesis.) The permutation of nucleotides in the genome is a blueprint for instructions governing every aspect of an organism's physical and mental being, so in fact this is a very useful representation to understand. It is no coincidence that the discrete nucleotide sequence of the genome has become synonymous with the genome itself. The transcriptome and proteome of an organism (comprised of its RNA and expressed proteins, respectively) also have similar discrete signal representations. RNA molecules are made up of discrete nucleotides like those in DNA, while proteins are

made up of amino acids.

Traditionally, deriving a complete, discrete sample representation of the genome has been a challenge due to its minuscule size and the problems associated with making such molecular-level measurements. Instead, for decades there have existed several technologies to determine shorter, partial representations of the genome by indirect means, i.e. by observing the hybridization of short sections of DNA strands with their complements, and then inferring the original DNA strand composition from it. Some such technologies are Southern blots, microarrays, and PCR. In fact, the Southern blot, named after its inventor, Edwin Southern, was so influential that it inspired later techniques named Northern, Eastern, Western and Southwestern blots for RNA and protein detection.

A sea-change occurred with the advent of the first sequencing method, Sanger sequencing. Today, next-generation sequencing (NGS) technology, buoyed by rapid advances in optics, nanotechnology and computer science, has made it possible to derive a complete, high-fidelity sampling of the genome. Instead of observing and measuring the whole-sequence hybridization of short DNA sections, we record *base-by-base* hybridization of DNA sections, using fast, redundant measurements, followed by a computational phase involving the alignment and assembly of those sections. The genomics world is at an exciting stage where several types of sequencing technologies have not only been developed, but are being further perfected every day while providing us with freely flowing genomic and transcriptomic data. For the first time, the genetic material of every organism has the potential to become a known quantity.

Unlike analog-to-digital signal converters, technologies to sense genetic material are expensive. Sequencing may be the gold standard for the rigorous assessment of

genomic material, but its cost is still extremely prohibitive for practical non-research use. Until the price of sequencing falls from bench-levels to bedside-levels, it is critical to focus on other "partial" sampling methods like microarrays and microfluidic devices that are readily available and less expensive for practical purposes.

Until then, we can take advantage of the fact that an organism's DNA is unvarying over its lifetime*, so it only needs to be sequenced once to be known. We can leverage our knowledge of fully-sequenced genomes to make informed pronouncements on the data that partial-sampling based technologies provide us.

The two technologies that we focus on re-engineering in this thesis are microarrays and molecular beacons. These are two mainstream and popular methods for DNA detection, but can also be modified for the detection of other biomolecules. In Chapter 5 we discuss sequencing technology, the current gold standard in DNA assessment that is growing in popularity but still has prohibitively high costs.

2.2.1 DNA Microarrays

The generic microarray refers to a device that is a solid surface with thousands of sensors attached – these sensors work in parallel to detect a target. The nature of these sensors will vary according to the target to be sensed; they may be nucleic acids, proteins, tissue, cells etc. In DNA microarrays, these sensors, or *probes*, are short strands of DNA to which the target DNA is expected to adhere as dictated by Watson-Crick base pairing [12]. Each *spot* on the microarray consists of multiple copies of a probe. For the purposes of this thesis, a probe refers to the collective of its copies.

*Epigenomics research targets the molecular and structural modifications to DNA that do not change the underlying sequence but ensure that the right genes are expressed at the right time [11].

The microarray hybridization process is as follows. First, the target DNA sample to be analyzed is fluorescently tagged before it is exposed to the microarray. A DNA subsequence in a target will tend to bind or hybridize with its complementary subsequence on a microarray to form a stable structure. The extraneous DNA is washed away so that only the bound DNA is left on the array. The array is then scanned using laser light of a wavelength designed to trigger fluorescence in the spots where binding has occurred. A specific pattern of array spots will fluoresce, which is then used to infer the DNA makeup of the test sample.

One of their major advantages when they were first invented was the high-throughput multiplexed nature that they brought to DNA sensing. Tens of thousands of probes were simultaneously exposed to the same target molecules, drastically speeding up experimental turnaround time. Consequently, microarrays have a variety of applications. In gene expression experiments, thousands of genes can be simultaneously monitored to study the effects of treatments, progression of diseases etc using the mRNA they produce. In alternative splicing[†] arrays, probes are designed to be specific to the expected splice sites of predicted exons.

Much of the work in this thesis centers around bacterial identification. Currently, microarrays used for this purpose depend on (multiple copies of several) unique gene identifiers in the target molecules, typically in the 16S or 23S housekeeping genes, to distinguish between different species. Therefore, it is not possible to create a single microarray with the capacity to detect the range of thousands of strains of multiple bacterial species that exist and have been sequenced, let alone those that are yet

[†]Alternative splicing: the same coding DNA segments, or exons, can splice together in different ways to form different proteins. Scientists are interested in studying which splice variant gives rise to which protein.

unknown.

Furthermore, the genes that contain the unique identifiers are only a tiny fraction of the full genome targets, so it is necessary to amplify their concentration exponentially before they are exposed to the microarray. Otherwise any binding will be overwhelmed by the remaining genome DNA. If the appropriate gene sections are amplified, there will be a sufficient number of target molecules to bind to probes, so we can visibly detect the variation in probe fluorescence that indicates target presence. This amplification step PCR (Polymerase Chain Reaction), is regarded as a standard pre-processing step to increase the concentrations of specific sections of DNA, and is used before most biomolecular techniques, including sequencing.

However, PCR amplification comes with several downsides. It contributes to added processing time, cost and often nonlinearly amplifies the DNA quantities. Nonlinear amplification ambiguates the original DNA concentrations in the sample, and as a result, a PCR-based result is not always assured to be representative of the original sample. CS-inspired genomic tools may offer alternative DNA sensing methods that obviate the need for PCR.

2.2.2 Molecular Beacons

Molecular beacons are free-floating single-stranded oligonucleotide hybridization probes that can report the presence of nucleic acids. They have a stem-and-loop structure, with a fluorophore (fluorescent dye molecule) that is activated in the presence of a target of interest, but otherwise quenched by a quencher molecule [13]. There are two advantages over microarrays from a biological perspective. One, their free-floating nature increases the probability of contact between probe and target. Two, they are useful in detection situations where it is either not possible or desirable to

isolate probe-target hybrids from the surplus probes.

Molecular beacons also hold practical advantages over microarray use. Since they are freely-flowing, they can be encapsulated in a microfluidic device, which is a significantly more convenient form for clinical use than a microarray. Furthermore, hybridization time with beacons is much shorter than the ~16 hours that a microarray requires in a hybridization oven for uniform exposure.

For a beacon to fluoresce in the presence of a target, it is important for the target-beacon binding to be more energetically favorable than the stem-stem binding within the beacon. This is usually facilitated if the beacon's loop sequence is as similar (specific) as possible to the target to be sensed, so that their binding is strong. A sloppy hybridization between the beacon and an unintended target molecule is not likely to be energetically stable enough to cause the stem to unzip and the fluorophore to be activated. Beacons are therefore especially specific DNA sensing tools.

2.2.3 Sequencing

Sequencing technology began with Automated Sanger sequencing, which involved several arduous steps, including a painstaking capillary electrophoresis step. It was revolutionary in its time, and is also the technology that delivered the human genome sequence to us, albeit over a period of almost 10 years. Today, sequencing is undergoing a revolution with the emergence of several radically different sequencing methods that grow faster and cheaper by the day – many promising to deliver $1000 genomes within the next few years. Some instruments that are already commercially available are from Roche, Illumina, Life and Pacific Biosciences [14].

The end-to-end sequencing of a DNA sample involves several steps that may be grouped into: (1) Template preparation, (2) Sequencing and imaging, (3) Data anal-

ysis. In the template preparation stage the DNA sample is amplified, fragmented into smaller pieces, and finally modified for the sequencing reaction, to generate a "template library" that is representative of the sample. These templates are then immobilized to a single solid surface in different ways depending on the instrument in question. This allows thousands to billions of sequencing reactions to be performed simultaneously on the same surface.

The actual sequencing process may currently be accomplished by a few different means, but the most popular by far is *sequencing by synthesis* where the complementary nucleic acid strand is synthesized by the base-by-base addition of nucleotides, along with a DNA polymerase enzyme that enables the synthesis. This is the process that the Illumina/Solexa GA_{II}, the machine whose sequencing data we analyze, follows. Each base has a dye molecule attached to it, whose fluorescence indicates the synthesis of another nucleotide in the complementary strand. The sequencing process consists of several *cycles* each of which involves the sequential addition of all four bases A, C, T and G. Sequence fluorescence is recorded after every base addition in every cycle, for the thousands to billions of sequencing processes that are happening on the same surface [14]. Therefore a 90-cycle sequencing process will ultimately yield thousands to billions of 90mer sequencing reads to be further analyzed. A read may also be referred to as a *kmer* if it is k bases long.

Finally in the data analysis stage, the series of fluorescent images that were recorded during sequencing undergo several signal processing operations like filtering, and are finally translated to 4-base reads by a 'base-calling' algorithm. If the purpose of the sequencing reaction was full genome reconstruction, they are then aligned to a reference genome if there is one, and assembled.

2.3 DNA Hybridization Models

Practically all present-day genomic sampling tools are based on complementary hybridization. The DNA molecule of interest binds with its complement; this binding is measured, and the original DNA molecule's nucleic acid composition is inferred from it. Tools like microarrays and beacons use the complementary binding between probe and target DNA strands, and then measure emitted fluorescence of each bound section. Therefore it is especially important to establish a model to predict the binding strength between a probe and the target is intended for, in order to design them appropriately.

Sequencing-by-synthesis technologies, as the vast majority of NGS technologies are, also depend on complementary hybridization. However they use *single base* binding and fluorescence to ascertain the DNA sequence one base at a time, so there is no probe design component and hence no critical dependence on a hybridization model. Single base binding is essentially binary – either it is bound and fluoresces, or not.

There are several factors that influence the hybridization affinity of a DNA fragment with another. We discuss two different approaches here: the first is computational, and the second, based on biochemistry and thermodynamics. The first is the Smith-Waterman alignment model [15]. It is based on the Smith-Waterman algorithm, a dynamic program that finds optimally local alignments between any two discrete sequences. The second is the SantaLucia Nucleic Acid Thermodynamic model, which is based on the empirical and theoretical predictions of the bound and unbound free energies of any pair of nucleotides [16, 17].

2.3.1 DNA Hybridization Affinities via Smith-Waterman Alignment

The Smith-Waterman alignment when given two discrete sequences as input will return the most similar local region between the two sequences. It compares segments of all possible lengths, calculates the corresponding sequence similarity according to a given scoring system, and outputs the optimal local alignment and similarity score. The cost model assigns costs to every match, deletion, substitution, and insertion in an alignment. When two nucleic acid sequences hybridize, it is based on the Watson-Crick rule of complementary base pairing. Therefore, to find the hybridization affinity between two sequences P and Q, we can first calculate the sequence similarity between a sequence P (or Q) and the reverse complement of sequence Q (or P). For example, if we have two sequences P = 5′-CCCTGGCT-3′ and Q = 5′-GTAAGGGA-3′, we first take the reverse complement of P = $AGCCAGGG$, and finds is optimal S-W alignment with Q:

$$\begin{array}{c} AGCCAGGG \\ \quad\;\; | \; | \; | \; | \\ GTAAGGGA \end{array}$$

where the regions of similarity in the two sequences have offsets of 5 and 4 respectively in order for them to be maximally aligned. In translating this to be biologically useful again, sequence P is once again reverse complemented to get 5′-CCCTGGCT-3′, and the SW alignment with 5′-GTAAGGGA-3′ is shown as:

$$\begin{array}{c} 3' - TCCC - 5' \\ \quad\;\; | \; | \; | \; | \\ 5' - AGGG - 3' \end{array}$$

Once we have an alignment for a given sequence pair, secondary parameters can be calculated to serve as metrics for hybridization affinity. There are several lists of

such parameters that are charged with predicting hybridization affinities; the most comprehensive of them cites 12 parameters [18].

Ideally we would like a hybridization affinity model that takes into account every sequence feature that could potentially affect hybridization. Here, in the absence of such a model, we use percent identity (defined as the percentage of aligned bases in the S-W alignment) as a measure of hybridization affinity. We also check each probe's secondary structure, to ensure that the probes do not fold on to themselves in a thermodynamically stable configuration [19]. Such folded structures may hinder a probe's hybridization with a target.

We recognize that sequence similarity is not the perfect model for spot intensity, especially since it does not incorporate critical biochemical factors that influence binding at a spot. In [20] we outline several post-S-W alignment parameters that we empirically estimate from experiments, that we determine to be significant in determining hybridization intensities.

2.3.2 DNA Hybridization Affinities via a Nucleic Acid Thermodynamics Model

Our goal is to estimate the probe-target hybridization intensities at each probe spot, given probe and target DNA sequences. In Section 3.4 we estimate DNA hybridization affinities by using the Smith-Waterman alignment algorithm. However, at their core, it is the free energies that are spent or gained that are responsible for the formation of a duplex between two molecules, or not. It therefore follows that we may use a thermodynamic model for DNA-DNA hybridization [16, 17], to give us a physically realistic model for hybridization affinities. For instance an A-T bond formation is a double covalent bond between hydrogen atoms, which releases relatively less free

energy since it is relatively less stable than a C-G bond, which is is a triple covalent bond between corresponding hydrogen atoms on each base. More complicated to quantify are the free energies when there are base mismatches.

Unlike many other models, this model, also incorporates thermodynamic parameters for mis-hybridizations between two DNA sequences (accredited to SantaLucia [16, 17]). The frequent occurrence of base mismatches in probe-target hybridizations during the probe design phase makes it critical that we estimate affinities in the face of these mis-hybridizations as accurately as possible. Using Smith-Waterman sequence alignments to do so is inexact for cases of mismatches, since the algorithm is highly dependent on the type of penalty that gaps are assigned by the alignment scoring matrix in use. There is no consensus on which penalty values most accurately reflect hybridization reality.

The SantaLucia thermodynamic model is commercially available from the company DNA Software, Inc. We purchased two component software packages: (1) ThermoBlast, which performs fast alignments of sequences against large genome databases to discover thermodynamically stable hybridizations, and (2) Visual OMP DE, which can simulate hybridization experiments with detailed solution conditions through scripts launched from the command line, and generate results for melting temperatures, Gibbs free energy (ΔG), and the percentage-based concentration of each resultant species post-experiment. There is also the capability to visually generate the secondary structure for each monomer, homodimer and heterodimer species formed from the constituent probes and target fragments.

We follow a two-step procedure in determining the hybridization intensities of a given set of probes against a given dictionary of potential targets. First, we run each probe against each double stranded target genome, using ThermoBlast, to discover

all possible target fragments at which there is significant alignment (thresholded at ΔG = -16kcal/mol). Second, we use Visual OMP DE to run a script to simulate each hybridization experiment between a single probe and a single target, using these target fragments to simulate the target. Every experiment simulation script contains information on: the probe sequence, the target fragments from ThermoBlast it was shown to hybridize with, and conditions for the experiment. Experimental conditions included probe and target concentrations, assay temperature, hybridization cocktail composition (Mg++, Glycerol, DMSO, Formamide, TMAC, Betaine concentrations), and pH. This procedure is repeated for each probe-target pair.

The hybridization results of each experiment in Visual OMP DE produces a data file, which is parsed for the concentrations of resultant species at the end of the hybridization. In each data file we use the percentage of probe-target heterodimer structures formed, i.e. the percentage of target fragments that are bound to probes, as an estimate for the hybridization affinity of that probe-target pair.

Chapter 3

Compressed Sensing Microarrays

In this chapter we introduce a new design of microarray, the Compressed Sensing Microarray (CSM). Its core design change is the use of combinatorially sensing probes – probes that each identify a group of targets, rather than a single target. After exposure to a sample, the pattern observed in the microarray can be decoded to reveal the presence and concentrations of targets in the sample using Compressed Sensing reconstruction algorithms.

This design allows a decrease in the number of probes necessary for sensing, but is still *specific* for a given set of targets, since the probes are carefully designed to hybridize with them. In Chapter 4, we describe a design for *nonspecific* sensing, where the probes are generated independent of any target set.

3.1 Concerns with Traditional DNA Microarrays

There are three issues with traditional DNA microarrays that stem from the fact that each sensing spot is designed to uniquely identify only one target of interest.

The first concern is that very often the targets in a test sample have similar base sequences, causing them to hybridize with the wrong probe. These cross-hybridization events lead to errors in the array readout.

The second concern is the restriction on the number of targets that can be identified. In a typical biosensing application, multiple organisms must be identified,

which necessitates a large number of spots. As a consequence readout systems for traditional DNA arrays are difficult to miniaturize.

The third concern is the inefficient utilization of the large number of array spots in traditional microarrays. While the number of potential agents in a sample can be very large, only a few agents are expected to be present in a significant concentration at a given time and location or in a given air/water/soil sample. Therefore, in a traditionally designed microarray only a small fraction of the large number of spots will be active at a given time, corresponding to the few targets present.

3.2 Designing Compressed Sensing Microarrays

To combat these issues, we propose a new microarray architecture using "combinatorial testing sensors" in order to reduce the number of sensor spots. We refer to this new type of array as a Compressed Sensing DNA Microarray (CSM), since it is based on the nascent theory of Compressed Sensing [20–23]. Each spot in a CSM identifies a *group* of target organisms, and several spots together generate a unique pattern identifier for a single target. Designing the probes that perform this combinatorial sensing is the essence of the microarray design process, and what we aim to describe in this section.

Let the $N \times 1$ vector x represent the concentrations of N possible organisms in a sample; we assume that only $K \ll N$ of them are actually present. We aim to design a microarray that implements a Φ that satisfies the RIP. Such a CSM will be able to reconstruct an estimate of x using $M \ll N$ probe spots.

To obtain a CS-type measurement scheme, we can choose each probe in a CSM to be a group identifier such that the readout of each probe is a probabilistic combination of all the targets in its group. The probabilities are representative of each probe's

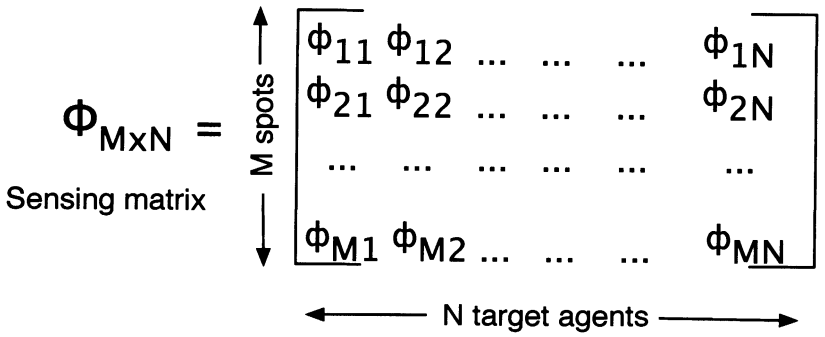


Figure 3.1 : Structure of sensing matrix Φ of a CSM with M spots identifying N targets

hybridization affinity (or stickiness) to those targets in its group; the targets that are not in its group have low affinity to the probe. Then the readout signal at each spot of the microarray is a linear combination of hybridization affinities between its probe sequence and each of the target agents.

Figure 3.1 illustrates the sensing process pictorially. To formalize, we assume there are M spots on the CSM and N targets; we have far fewer spots than target agents, so that $M \ll N$. For $1 \leq i \leq M$ and $1 \leq j \leq N$, the probe at spot i hybridizes with target j with affinity $\phi_{i,j}$. The target j occurs in the tested DNA sample with concentration x_j. Then the measured microarray signal intensity vector $y = \{y_i\}$, $i = 1, \ldots, M$ fits nicely into the basic CS measurement model of equation 2.1. Each $y_i = \sum_{i,j} \phi_{i,j} x_j$.

In related work, group testing has previously been proposed for microarrays [24]. The chief advantage of a CS-based approach over direct group testing [25] is its information scalability. With a reduced number of measurements, we are able to not

just detect, but also *estimate* the target signal. This is important, because often pathogens in the environment are only harmful to us in large concentrations. In the language of CS, the Φ that is used in group testing is binary. In CS, the measurements y are linear sums of the elements of x, with weights prescribed by the corresponding row of Φ. In group testing, each measurement in y is a superposition or "or" operation over the elements of x, with the corresponding row of Φ determining whether or not each element contributes to that superposition.

In the following sections we describe the steps required to design a CSM. In brief, the step-by-step process is:

- Decide on target dictionary (x)
- Decide on an appropriate CS measurement matrix (Φ)
- Use an appropriate algorithm to assign target to columns of Φ, and design probes for each row of Φ
- Calculate the precise DNA hybridization affinities of each probe-target pair to give each element of Φ
- Once an experiment is performed, input y (probe readout) and Φ to a CS reconstruction algorithm, to decode x

3.3 Target Assignment and Probe Selection Algorithm

Given a target dictionary x, our goal is to design a microarray that implements a CS measurement matrix Φ with the RIP. In a CSM we make the design assumption that we can choose the Φ to achieve in the microarray. This specifies both the number of probes M and the desired degree of hybridization $\phi_{i,j}$ between probe i and each

potential target j. We choose a binary Φ, whose entries are given by a Bernoulli distribution (for example), which satisfies the RIP. This will facilitate the probe design process, since we only need to find probes that satisfy a 1 or 0 hybridization affinity instead of a real value – which amounts to maximizing or minimizing the hybridization affinity of a probe with a target. Our design goal is to assign probes to Φ one row at a time, by finding DNA sequences that are shared between groups of targets, while simultaneously assigning targets to the columns of Φ. We tackle this problem by first assigning targets to columns, and then designing the probes for each target group.

A convenient guideline for target assignment to columns would be an existing grouping of targets based on their shared sequence similarities. We use such a grouping, from the COGs (Clusters of Orthologous Groups of Proteins) database [26], an NIH-governed database that organizes prokaryote and unicellular eukaryotes into groups based on the similarity of their protein sequences. Since protein sequences can be translated back to DNA, this gives us a basis for grouping organisms according to their DNA sequence similarity. The advantage of using the COGs database is that it is based on exhaustive alignments between organisms and is therefore a good leapfrog into target grouping. Furthermore, as more genomes are sequenced, they are added into the grouping so that the database is continually expanded. Currently there exist 4872 COG groups, to which 66 microbial genomes (bacteria, but also species from Archae) belong. On average each species belongs to $\sim$12 COGs.

Given a set of targets, we find all COGs they belong to. Then, starting with the first row of Φ, the COG whose target grouping best approximates that row of Φ is assigned to it. This sets the target assignment to columns. Fixing this target assignment, COGs are identified to approximate the remaining rows. Note that each subsequent COG assignment after the first is more and more constrained. This is

repeated for each row of Φ, in order to maximize the number of COGs that readily conform to the row weights of Φ. The optimal target assignment is the one that maximizes the number of rows of Φ that are matched by COGs.

Once targets are assigned to columns of Φ, we are left with the task of designing probes that produce target hybridization affinities as dictated by the row weights of Φ. In theory, this should be a relatively straightforward process, since the targets were grouped according to COGs, which in turn are grouped according to sequence similarity. However, in practice, probe selection is a delicate task due to the finicky nature of DNA hybridization.

We begin by choosing an appropriate probe length for the CSM; for an oligonucleotide microarray this can be between 20-70 bases. We also specify that all probes have GC-content between 40-60%, so that their melting temperatures are uniform. The target genomes themselves are available from the NIH NCBI website*, there were a total of 2277 bacterial chromosomes sequenced, from 1435 different species. For the target sequences we use downloaded sequenced chromosomal† DNA of bacteria from the NCBI website. There are several procedures to select probes. Here we show one simple algorithm for probe selection; for a variant, we refer to [22].

3.4 Using Smith-Waterman Alignment to predict Φ

Given a set of probes and targets, we need to predict their DNA hybridization affinities to determine Φ as precisely as possible. Accurate decoding of the reconstruction

*http://ncbi.nlm.nih.gov/genomes; viewed on April 27 2010

†Plasmid DNA may also be similarly used but is not included here. Plasmid DNA is typically shorter than chromosomal DNA, 1-1000 kilobase pairs, so may need differently sized probes for identification.

algorithm depends on being able to predict Φ as precisely as possible. An element of Φ should reflect the *spot intensity* of a probe-target hybridization, since the measurements available to us are in fact the spot intensities.

There are two main steps to translating (probe, target) $\rightarrow$ spot intensity. First, we require a multivariate model that uses features of the probe and target sequences to predict a hybridization affinity value between them. Second, we must translate this hybridization value to a microarray spot intensity using a model that takes into account physical parameters of the experiment such as background noise, saturation effects, etc. In Section 3.5 we look at one possible spot intensity model, and try to incorporate its potential effects in the decoding algorithm. However, for the rest of this thesis, we proceed under the assumption that hybridization affinity is equivalent to spot intensity. In the case of the practical experimental design of a CSM, there would be a calibration step where the fluorescent intensity in RFU's was recorded for every bacteria-probe pairing, for fixed concentration levels for both.

3.5 Decoding Nonlinear Measurements via Belief Propagation

We have already established sparsity in x – the number of pathogens likely to be present in any given sample is typically small compared to the total number of pathogens in our dictionary. We recognize a second type of sparsity here. The biology of COGs tells us that the number of biological agents that *share* similar DNA fingerprints is likely to be small compared to the total number of agents; i.e. common genetic material is sparse [27, 28]. This situation translates to a sparsity in hybridization affinities between genes in target agents, and finally a sparse Φ matrix.

After the experiment is performed using an appropriate CSM, we use CS decoding algorithms to decipher the pattern present. The inputs to the algorithm are y, the pattern itself, and Φ, the matrix of hybridization intensities. A sparse Φ enables us to use decoding algorithms such as Belief Propagation (BP) [9] to reconstruct the target signal. Besides being fast, BP has the advantage of being easily adaptable to different signal models. We leverage this flexibility in adapting the BP algorithm to modeling nonlinearities and noise in measurement values.

The model that we use for spot intensity is the Langmuir Isotherm Model, a nonlinear concentration-intensity model, which relates the coverage of molecules on a solid surface to the concentration it exists in, at a fixed temperature. Its governing equation is:

$$y' = \frac{\alpha y}{y + \beta} + n \tag{3.1}$$

Here y' is the spot intensity, y is the concentration of probe-target hybrid molecules, α and β represent probe/target specific constants of hybridization and n is the noise component at the spot. If we abstract the nonlinearity as $T(\cdot)$, and the linear combination of gene concentrations as $L[\cdot]$, we can represent the kth spot's intensity as:

$$y_k = T(L[x_1 ... x_n]) + \mathcal{N}(\mu, \sigma^2), \tag{3.2}$$

Here again we assume the measurement noise to be gaussian. Furthermore, the nonlinearity $T(\cdot)$ does not have a well-defined inverse; as concentration increases, the intensity measurements recorded will plateau off at α. This nonlinearity in the microarray measurements needs to be addressed during the BP decoding process.

Algorithmically, from the perspective of CS decoding, this means that instead of the true measurements $y = L[x]$, we are supplied a nonlinear, noisy version, $y' =$

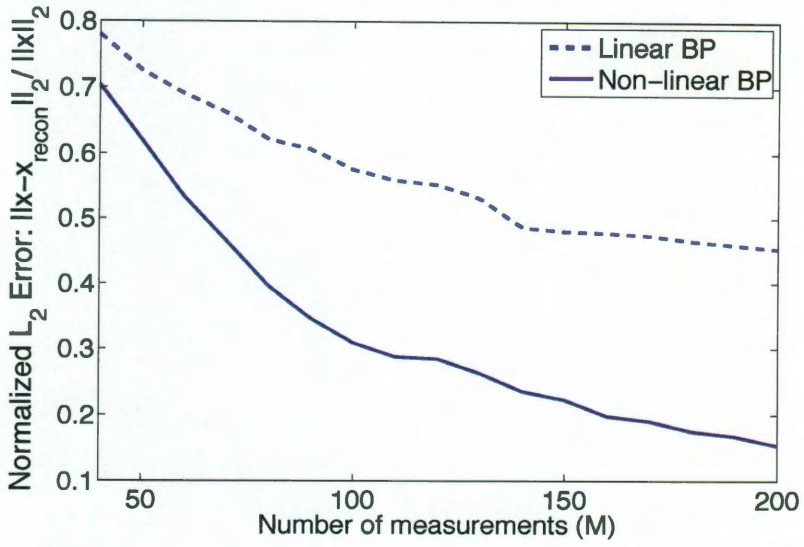


Figure 3.2 : Plot of normalized L_2 measurement error vs. number of measurements comparing decoding nonlinear measurements by our modified BP version and BP that ignores the nonlinearity. Number of signal coeffs $N = 200$; $\alpha = \beta = 25$; $\sigma_y = 2$

$T(L[x])$. We modify the original BP algorithm by adding variable nodes for the nonlinearity $T(.)$ and noise, $N(i)$. For further details on the algorithm and numerical results, we refer to our work in [21]. As an illustration of the algorithm's working, Figure 3.2 demonstrates the change in L_2 reconstruction error of the true signal (sparsity 10%) against the number of measurements (i.e. DNA spots), using our nonlinearly modified BP algorithm as well as the regular BP decoding algorithm that ignores the nonlinearity. We notice that by taking into account the nonlinearity and reversing it during the decoding process as our algorithm does (while throwing away the saturated measurements), the L_2 decoding error converges to a smaller value than if we had ignored it.

3.6 Problems with COGs-based Probe Design

The probe design process outlined above involves first choosing a Φ appropriate for Compressed Sensing, and then picking the probes whose hybridization intensities fit that Φ, using the COG-grouping of targets as a constraint. There is one major problem with this approach – the use of COG-grouping of targets as an indicator of similarity in targets' probe hybridization intensities. COGs, while a starting point for probe design, are largely a functional grouping based on similarity in protein structure. There is certainly a biological mapping between DNA sequence and protein sequence, but it is not 1-to-1. Ultimately, COGs-based target sequence alignment is a suboptimal probe selection process; there exists a better target grouping based on true DNA sequence similarity that would give more closely aligned targets and therefore better probe representatives. Algorithmically, the procedure of finding probes whose target hybridization affinities conform to a Φ, even approximately, is tedious both in terms of the repeated multiple sequence alignments necessary for probe selection, and the target assignment process via COGs is also cumbersome.

It is also useful to note that CSM's do not offer the possibility to obviate the need for PCR. This is because each CSM probe is specifically designed to hybridize with a single section of each target genome in its group. This implicitly forces the need for PCR amplification for these specific sections in each different target genome. Furthermore our decrease in number of probes may be compensated for by the increase in number of primers needed. We see in Chapter 4 that a different design of combinatorial probes offers the promise to eliminate PCR, in addition to other new advantages.

Algorithm 1 : Probe design for CSMs

1. For rows i :1$\rightarrow M$:
   (a) Use a multiple sequence alignment algorithm between the targets with positive weight in that row
   (b) In regions of similarity, use a sliding window of probe length to find all potential probe candidates
   (c) Filter these probe candidates to keep the ones that satisfy the GC-content constraint
   (d) Calculate precise hybridization affinities between all remaining probe candidates and the targets in that row
   (e) If the hybridization affinities of a probe with all L targets are close to their corresponding $\phi_{i,j}$ values, store the probe;
2. Check for loop formation in the secondary structure of the *complements* of all the surviving probe candidates.
3. Choose the probe with zero or fewest loops. If more than one, choose the probe with the shortest length loop and highest target hybridization affinities.
4. End

Chapter 4

Nonspecific Sensing via Random Probes

Many molecular detection problems make the difference in our literal survival as individuals, communities, even as a species. Detection purposes are varied. We may want to detect objects that are harmful to us, such as environmental toxins, viruses and pathogenic bacteria. We may also want to detect others that are vital to us, such as oxygen levels or blood glucose for diabetic patients. Finally, we would like to detect and characterize unknown quantities, because they *may prove* harmful or vital to our existence, but are as of yet unidentified.

4.1 Need for nonspecificity in bacterial detection

The resolution of every detection or classification system is limited to the scope of the sensors it contains. Therefore it is vital that a detection system's sensors cover the full range of potential targets in order for it to be useful. Fortunately, in most cases this range of detection targets is known in advance. For instance, in the case of glucose monitoring or many infectious diseases caused by pathogens, the characteristics of the specific molecule or organism of interest are well-studied (the structure of glucose or the characteristic spores of Anthrax), and we can use those same characteristics in sensors for their detection.

However, there is also a practical need for *nonspecific* sensors that can identify new molecules and organisms. This is particularly exacerbated for the situation of

bacterial detection. Studies of prokaryotic diversity estimate that given how long bacteria have existed on the planet, there are tens of millions of species and strains of species that exist but are yet unknown to us. Given that the current number of sequenced bacterial genomes is less than 3000, the means to identify and characterize these unknown species is essential. Furthermore, many microbial species are constantly mutating through the recombination of DNA between one another, from their host, or from bacteriophages (viruses). These fast-moving mutations may have effects that range from the benign to harmful; one less threatening example we see annually is the new strains of the flu virus that develop.

Besides the vast spread of bacterial species that occur naturally in the world, the development of the fields of synthetic biology and genetic engineering have given humankind the power to create "new" species. While these fields are developing quickly and are typically only used toward the gain of society, health and medicine, there exists the danger that these techniques may be used nefariously as well, namely to create new bacterial biological warfare. History is rife with examples of such. As early as 1200 B.C., the Hittites drove bubonic plague-sufferers into enemy lands to infect them. Anthrax and other harmful species have also wreaked havoc as biological weapons for a long time; one of Anthrax's first documented uses was in the 1930's when the Japanese used it on prisoners of war in Manchuria. Given the potentiality for newly engineered biological species to be used for such ill-desired purposes, it is critical to develop the means for their identification, or at least characterization.

Currently used bacterial detection systems are all purpose-built. They use specific sensors to detect characteristic compounds produced by those species; these compounds may be DNA, proteins or other biological molecules, and are identified by their characteristic response to the sensors used to detect them. For instance, the

clinical "rapid" test for strep throat uses an antibody that reacts with a specific carbohydrate antigen in the cell wall of *Streptococcus pyogenes*, the species responsible for the infection (but this test is not always accurate). Even today, the clinical standard in most bacterial detection tests is culturing – the growth of the bacterial species and then its visual inspection for growth patterns or identification at the cell-level.

However, for the case of unknown mutated or engineered species, only reading a genome will tell us that it differs from all previous organisms in our library. Sequencing does exactly that, and it remains the holy grail for genomic species identification. Currently there is no other technology platform that allows us to identify new species as well as the old ones on the same platform. But the cost of sequencing technology is still exorbitant, and it will be many years before it makes its way to either medical microbiology labs or the battlefield.

4.1.1 Genome-based bacterial detection

In current genome-based bacterial identification methods, the focus is on the 16S and 23S ribosomal genes – two specific bacterial genes that are highly conserved in bacteria and archaea and therefore have several sites that are shared between them, yet have other sites with enough variation to pinpoint many bacteria at the species level. The chief problem with using the 16S and 23S ribosomal genes is that they are extremely short – only 1542 nucleotides and 2904 nucleotides long respectively. This means that in order to identify them, they must undergo a serious amplification step by PCR to increase the number of their copies after the genomes have been isolated from the bacterial sample. After PCR the amplified gene fragments are subject to a molecular detection stage, where a microarray, Real-Time PCR, or other molecular biology means are used to identify the variation within them. This is done

by complementary DNA hybridization through the use of DNA probes – fluorescently-modified DNA fragments that are complementary to the unique 16S or 23S identifiers. More recently, it has become more common to sequence the 16S or 23S gene after PCR.

Genome-based detection methods for bacteria are not widely used, not even in clinical/medical labs – one of the more critical situations for bacterial identification in everyday life. There are several reasons for this. The foremost among them is cost. The reagents and instruments needed for PCR and the actual molecular detection technique are very expensive. Second, the time needed for this procedure is lengthy, and requires several specific, fragile protocols to be followed precisely in order to get sensitive results. Finally, any molecular detection device that is used will only contain sensors catering to a narrow range of bacteria, since each bacterium requires several 16S or 23S DNA probes unique to itself. This also makes it cumbersome and expensive to store many different devices with different probes in order to detect a large number of bacterial strains.

It is unfortunate that genome-based identification methods are not yet practical in situations outside the research laboratory, since they are the most conclusive means of species identification – whether known or unknown – due to the irrefutable evidence they provide. Even a visual inspection of two identical cell cultures may belie the fact that they are from two different, but perhaps closely-related species.

Realizing these challenges and the criticality of the goal at hand, we seek other means for the genome-based identification of bacterial species. Our proposed solution is to stay within the framework of existing molecular detection devices, but change the way in which the probes they use are generated. The guiding principle in this is the mathematics of Compressed Sensing, which confers several advantages on sensing

situations where measurements are taken in accordance with it. Interpreted for a genomic molecular sensing device, chief among these advantages are the ability to detect both known and unknown bacterial species, the use of far fewer probes for high detection accuracy, and the promise of eliminating PCR – all of which translate to lower costs and a faster detection process.

4.1.2 Random probes powered by Compressed Sensing

Compressed Sensing (CS) is a recent theory of signal processing which describes, as its name suggests, the sensing of signals in a special, "compressed" fashion in their subsequent reconstruction. The prerequisite to the application of CS is that the signals to be sensed either themselves be sparse, or be linearly transformable to some other sparse representation. In such cases, logarithmically fewer measurements are needed to reconstruct those sparse signals than previously dictated by signal processing principles.

The bacterial detection problem can be recast as a signal reconstruction problem if each element of the signal to be reconstructed by CS is allowed to correspond to the concentration of a bacterial species to be detected. Furthermore, in most environmental or medical samples the number of bacterial species that are present in a significant concentration is few compared to the total number of bacterial species to be detected, so the sparse detection problem is also a sparse reconstruction problem in the canonical basis.

The CS measurements needed for sparse signal reconstruction cannot be arbitrary, and CS theory tells us that they must be taken in a *holistic*, yet *differential* fashion – each sensor takes measurements of the signal as a whole, but does not weight every part of it the same way. Together, enough sensors capture the signal diversity needed

for its reconstruction. There is great flexibility in how differential each sensor must be; in fact, one way to ensure that measurements of the signal are adequate for CS is if they are taken randomly. For instance, a random iid gaussian measurement system where each sensor randomly weights parts of a whole sparse signal satisfies the holistic-yet-differential sensing property, but if every sensor in the system only weighted the (say) first element of a sparse signal, it would not. Formally, this property is known as the Restricted Isometry Property, and is described in detail in the literature [29].

In genome-based detection methods, each probe measures every bacterial genome in the set of interest through complementary hybridization. To satisfy CS requirements, each probe must measure genome features that differentiate some of them from others; then a set of sensors together captures enough of the diversity of the entire genome set for their individual identification. Since these measurements must also be holistic over the whole signal – in this case the entire genome set – they cannot all be unique to each genome. For instance, unique 16S-based DNA identifiers for bacteria will only measure the bacteria they occur in, so do not qualify as appropriate CS-based sensors. Instead, we may choose specific DNA sections that match in several different locations in many bacterial genomes. This way, each DNA probe can holistically but differentially measure the full set of bacterial genomes. This is the idea behind the combinatorial sensing of the CSM described in Chapter 3, where bacteria are grouped together functionally, and then shared DNA fragments are systematically extracted from each group by a multiple sequence alignment process.

However, combinatorial DNA probes still do not harness the full flexibility bestowed by CS sensing principles: that the method of taking measurements does not have many constraints, and in fact random measurements also satisfy the RIP required of CS. They are also specific to each group of targets, and cannot detect

any unknown, mutated or newly engineered species. This suggests the use of probes that are *randomly generated*, instead of specifically designed. Indeed, random probes generated independently of any bacterial target would also be nonspecific and allow us to identify and quantify unknown species. The question remains: can randomly generated probes also generate random-enough measurements that satisfy CS requirements? This is the answer we seek empirically in designing random probes for bacterial detection. Since this is in fact not a thought experiment but for actual detection, we ground our work in reality by creating random probes specifically for the DNA hybridization microarray. In this incarnation, we refer to it as a Random Probe Microarray (RPM). We describe how the detection abilities of such random DNA fragments fit the model for CS measurements in the remainder of this chapter.

4.2 Random Probe Microarrays

In a departure from the convention of careful probe selection algorithms, including those for the CSM, we suggest *random* probe selection as a means of populating a microarray. The design of this microarray enables *universal* sensing – the same set of probes may be used to sense any organism. Furthermore, this is done with a number of probes that varies only logarithmically with the number of organisms to be detected.

The working of the RPM is described by the CS-based equation 2.1, where each element of Φ represents the affinity of the random probe corresponding to the row it is in with the bacterium corresponding to the column it is in. From this angle, the setup is the same as that described for the CSM in Chapter 3.

The key difference in the RPM from the CSM is that the probes are generated randomly, so there is no preconceived Φ. We generate probes first, and calculate

the Φ they represent later. The choice we make to generate a large population of random probes gives us greater flexibility to impose thermodynamic stability and other biochemical criteria that determine probe quality. This is in contrast to the careful probe design process in the CSM, where we endeavor to find probes that satisfy a desired Φ. In using random probes, we leverage an important flexibility that Compressed Sensing theory affords us: random matrices preserve the information in sparse vectors with high probability. We bank on the possibility that random probes are likely to correspond to random enough matrices (Φ's), which in turn will automatically satisfy the RIP condition required of matrices for CS measurement. From our simulations we empirically observe that this is true, and that the Φ's of the RPM are good enough for CS measurements.

We impose only two parameters in the generation of random probes – randomly generated sequences of A, C, G, T bases – length and GC-content. But the probes thus generated are also subject to biochemical constraints required of microarray probes, such as uniformity in melting temperature and secondary structure avoidance. It is important that the melting temperatures of all the probes are similar to each other so that a single experimental protocol is valid for all of them. It is also important that their melting temperates are sufficiently high that they are close to the denaturation temperature for double-stranded DNA. This way during the cooling phase after denaturation, DNA strands first anneal with the random probe fragments instead of their complementary fragments. The other main biochemistry constraint imposed is secondary structure avoidance, which is important so that the probes do not fold into a hairpin or other structure themselves.

The DNA hybridization model that we utilize for random probe-target affinity prediction is based on the thermodynamics of nucleic acid, and is described in Sec-

tion 2.3.2. We choose this model instead of the Smith-Waterman alignment model primarily because of the need for accurate predictions of DNA binding in the context of mismatches. There could potentially be many of these due to the random fashion in which the probes are generated, as compared to the conventional tailored probe design process. The model we use estimates the probe-target affinity value element-by-element, for fixed probe and target concentrations. Each probe is far in excess of the average amount of target available to it, so we may safely add multiple target interactions with the same probe by assuming them to be independent. In a microarray each probe is fixed, also limiting interaction and competition between different probes. (We will see in Section 4.5.2 that in molecular beacons this assumption is even more grounded, since the beacons are in physically separate wells, and never interact with one another.)

While the affinity values are indicative of how many probes and targets are likely to form heterodimers, ultimately it is the fluorescent intensities from the fluorophores on the targets that are the measurements we use. There are several intensity models that describe the nonlinear behavior with target concentration; but these only come into play when target concentration is in excess of that of the probes. Therefore we remain in a linear regime, and our measurement intensities are indeed linear combinations of the bacterial weights. In real experiments, raw measurement values are always calibrated before interpretation. For hybridization calibration, the positive control used is the perfect complement of the probe in question. Experimentally, calibration is critical, and Section 4.5 describes how experiments may be used to estimate these affinity values instead of a hybridization model, inclusive of positive and negative control values. Therefore, the affinity values determined by the thermodynamics model are a good estimate for the true affinities that are in turn, strongly correlated

with the fluorescent intensities we see.

Using the notation defined earlier, our goal is to best recover x (which describes which targets are present and in what concentrations) from the experimental y values (random probe measurements of an RPM) and the predictions for the elements of Φ (affinities between the probes and targets). The equations these quantities fit into is the same CS equation in Section 2.1. We summarize the end-to-end design and operation of an RPM:

- Generate random probes: choose random probes of desired length and satisfying biochemical constraints for the microarray
- Define dictionary of targets to be detected
- Determine effective sensing matrix Φ: calculate using a DNA hybridization model to obtain probe-target affinities, or from calibrating through actual experiments;
- Perform experiment: probes and targets allowed to hybridize; hybridized spots fluoresce
- Decoding: decode measurements with sparsity-based (CS) methods using Φ to decipher targets

4.3 Implications of a Nonspecific RPM

The RPM is nonspecific by design, and inspires a paradigm shift toward nonspecific sensing in situations where the quantities to be sensed are unknown. In contrast, both traditional DNA microarrays and CSM's are target-specific. Once a probe set is generated corresponding to a desired target set, it is *only* sufficient for detecting

that target set. (Note that while the CSM uses non-unique probes to identify groups of targets, those probes are still only useful in identifying those specific groups of targets.) On the other hand RPM probes are not purpose-built for any specific DNA sequence, which enables them to detect *any* set of desired targets. The minimum number of probes required to detect a given target is still prescribed by CS theory, so an RPM should require just as few probes as the CSM. On the decoding side, we may be able to use many more CS data recovery methods, since our Φ is not necessarily sparse by construction as it was for the COGs-based design.

On a practical level, RPM's advocate the use of *genomic*-based sensing, instead of genetic sensing. By this, we mean that we use the entire genome of each target organism, and sense it holistically instead of amplifying and detecting only specific gene variants that are unique to each organism. It is our hope that full-genome sensing will obviate the need for PCR-based amplification, reducing the cost and time needed for pathogen detection.

The chief novelty of the RPM lies in its random probe design, which produces its universality and future proofedness. It also maintains some of the benefits that the CSM offers: fewer probes and a method to both detect target presence and estimate its concentration. A summary of some of the RPM advantages follows.

- Future-proofness: Random probes are future-proof in the sense that as more organisms are sequenced we only need to update our software (by determining their affinities with our random probes) and are still able to use previously manufactured microarray hardware for organism identification. Therefore a new organism is accommodated not by physically adding probes but by an additional column in Φ. Furthermore, the minimum number of extra probes needed for identification via CS methods grows logarithmically slower than the

number of new organisms added to the database/dictionary.

- Broad range applicability: the same microarray with randomly generated probes can be used to detect any desired target set; the only changes to be made are in the software on the decoding side. For example, in our application we have focused on microarray design for *all* bacterial species, enforcing the universal quality of the RPM.

- New species discovery: The RPM, being as broadly designed as it is, could alert us to new mutant, unsequenced or newly engineered species if an unrecognizable probe pattern is generated.

- Robustness to error: From a practical viewpoint, the random base substitutions that plague traditional probe manufacture are no longer a concern during random probe manufacture as each random probe is as good as the other. As long as the errors are known, the change can be incorporated in calculating Φ and the erroneously produced array can still be used.

- Simple probe generation: A fast and simple randomization routine for each probe replaces the currently arduous probe design process in a traditional microarray or CSM. Traditional painstaking probe design methods are superseded by post-experimental computation, which is cheap.

- Phylogenetic ancestry discovery: We envision that such a broadly designed microarray may also be used to illuminate new phylogenetic or evolutionary relationships between organisms by comparing the probe-affinity profiles they generate. This follows from the fact that a randomly generated probe set is able to cover greater spans of an organism's genome than the traditional probe

set which is generated from the section of an organism's genome that is unique to itself.

- Potential to obviate PCR: PCR machines are both expensive and time consuming in the process of DNA-based detection methods. In utilizing the *entire* genome of each bacteria, there may be enough locations with enough binding intensity such that there is no need to amplify any single region of the genome where unique species identifiers reside.

4.4 Simulations in silico: Bacterial Detection by Random Probes

We investigate the minimum number of critical parameters needed in the generation of random probes: probe length and GC-content, while subjecting them to biochemical constraints such as uniform melting temperatures and secondary structure avoidance. Probe sequences are picked from a random distribution (e.g., uniform) over the {A,C,T,G} bases, but the lengths they are determine their affinity for each organism. Longer probes are less likely to stick to a target's DNA, while shorter probes may stick so much that they complicate the decoding process. Therefore, picking the appropriate probe length can be linked to the target set of the microarray, in that the requisite probe length may vary according to the genome length of the organisms to be sensed. In practice probes may be of a length that is pre-determined to be best for the target organism set (either through simulation or experiment). In the application of the RPM to a bacterial target set we found that probe lengths between 19–23 result in good detection accuracy for a variety of target bacteria sets. Thus an RPM array can be manufactured in advance, before the target dictionary is even decided. The

GC-content of the probes that we select is also of importance. Subject to biochemical constraints, probes with higher GC-contents around 50-55% have the high melting temperatures that we desire. However, these probes do not have the same affinity for all organisms. Species that are AT-rich and have short genome lengths have poor affinity for them.

4.4.1 Hybridization affinity generation for Φ

We first choose a target set of 100 random bacteria, and obtain their complete genomes from the NCBI database. Next we create the probe set to be tested by randomly generating DNA probes by specifying a length and GC content within a narrow range of 50-55% for high, uniform melting temperatures. The thermodynamic affinity model calculates secondary structure, so we do not need to impose this constraint separately. In our case we generated several sets of random probes with lengths between 20-26 (typical lengths for oligonucleotide arrays). Probes and targets were specified to be at $10^{-6}M$ (call it P) and $10^{-14}M$ (call it T) molarity concentrations respectively. This target concentration value approximately represents the concentration of a real bacterial sample that we may use; $1\mu g/\mu L$ of E. coli is approximately $10^{-10}M$. The probe concentration is always in excess of the limited target DNA that is isolated from the sample. This is also important experimentally so that all target DNA fragments have several orders of magnitude more probes available to them during the cooling process after denaturation.

Simulations were run to compute the probe-target affinities in Φ using the nucleic acid thermodynamics model described in Section 2.3.2. Each simulation consisted of a single probe interacting with all target fragments from a single target, and determines the probe-target binding affinity for a single element of Φ. This is also a realistic

model experimentally for a microarray experiment, where the probes are anchored to a surface but are exposed to all different target fragments. Therefore, we do not need to consider the interactions of probes between themselves. In this way, the probe intensity recorded as an element of y truly reflects a weighted linear combination of its affinities for individual target molecules. Furthermore, nonlinear models like the Langmuir Isotherm do not apply here since our probes are in excessive concentration of our targets. Each target fragment has its choice of probe due to the flooding of excessive number of probes. After all the hybridization affinity simulations between the full set of probes and full set of targets, the Φ is created. Any scaling performed on the columns (rows) of Φ is mathematically equivalent to post-multiplying (pre-multiplying) the Φ matrix with an identity matrix with column (row) weights along its diagonal, and is thus only a scale factor on our species vector (measurements). However, it is only through experimental calibration of multiple probe-target pairs that we can determine if or what kind of scaling factor or threshold factor to use in our simulations. We refrain from using a scaling factor in our simulations since a realistic one can only be obtained experimentally. We do however use a threshold of 20%. We use the same Φ to both create the measurements and for decoding. These steps are described in more detail below. Note that we must normalize all columns before using it for decoding with our recovery algorithm, CoSaMP in this case. In section 4.6.2 we consider the effect of errors in our Φ by modeling them as perturbations.

4.4.2 Trends in bacterial detection using CS reconstruction algorithms

Once Φ is created for a set of probes and targets, we run a CS reconstruction simulation to see how well that set of probes detects the set of targets. Since everything is in silico, it is important to introduce realistic errors to represent a real detection

problem. We make three modifications. One, we threshold low values of Φ since these represent probes with low affinity for corresponding targets, and may not bind in a real experiment, or have extremely low fluorescence. Two, we convert all elements of Φ to be binary valued. This is an extreme step, which may be experimentally interpreted as a very coarse calibration for both Φ and probe measurement values $\mathbf{y}$ – instead of accurately measuring the true fluorescences of probes in an experiment, we simply claim that they either fluoresce or do not. This is actually the group testing version of the CS reconstruction problem. Three, we also add gaussian noise to the measurement vector y generated from $\Phi \times x$, typically of SNR 5-10, meaning that the signal energy is between 3-10 times as large as the negative control. If detection works for such coarse modifications, even at the expense of twice as many probes as targets, we believe that it will also work for an accurate Φ model, low negative control and high quality fluorometer. (Simulations of such accurate conditions confirm that the number of probes needed for detection drops dramatically for a real-valued Φ compared to binary, thresholded at 20% and SNR = 10.)

Overall, we saw that probe length 19-23 all have good detection rates for the vast majority of targets. However, there were a few that fared less well, as we see from Figures 4.1, 4.2 and 4.3. A closer look reveals that the targets that were more prone to being missed for a given set of random probes were the ones with lower GC-contents and smaller genomes. In the figures shown, the two targets with poor performance are Ehrlichia ruminantium and Onions yellow, both with GC contents below 30% and genomes approximately a million base pairs long. We see that for these targets, detection capability depends on the probe lengths. Length 19 shows good detection, and a low false positive rate, compared to 21 which does worse both in detection and false positives. Finally length 23 shows a slight improvement in detection than 21.

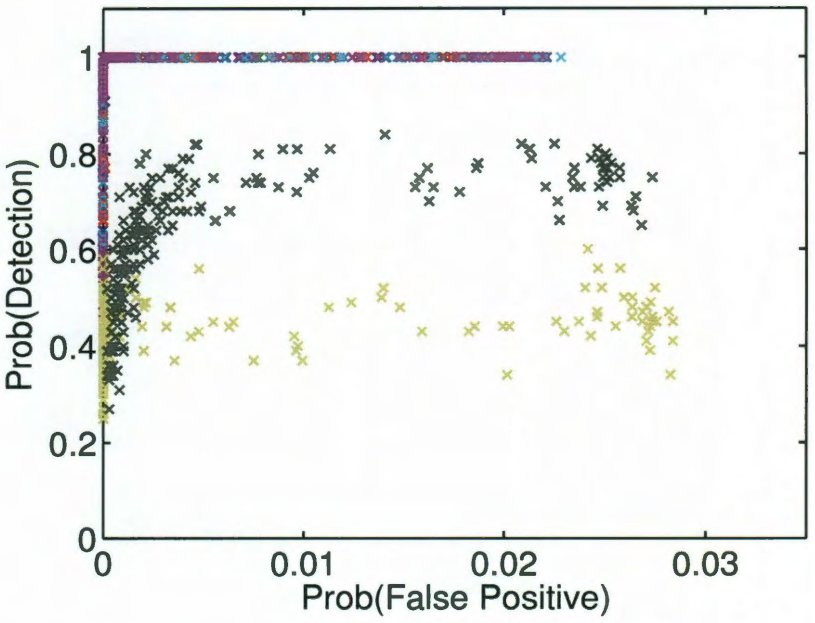


Figure 4.1 : ROC curve showing Probability(detection) vs. Probability(false positive) for random probes of length 19. The two curves in yellow and black are for the species Onions yellow and Ehrlichia ruminantium, which have small genomes and low GC content. Onions yellow performs even worse than E. ruminantium due to its smaller genome.

This trend can be explained by the fact that the shorter the probes are, the more likely it is that they have a binding site even in a shorter genome. However, after length 23, false positive rate is worse because of the sloppy binding that may occur between longer probes and targets.

This trend would imply that finding probes that are shorter than 19 is a step in the right direction for wider detection ranges. It is here that our restrictions on probe length and GC content come in; higher probe lengths and GC content imply higher melting temperatures, since more bonds are formed in a probe-target hybrid,

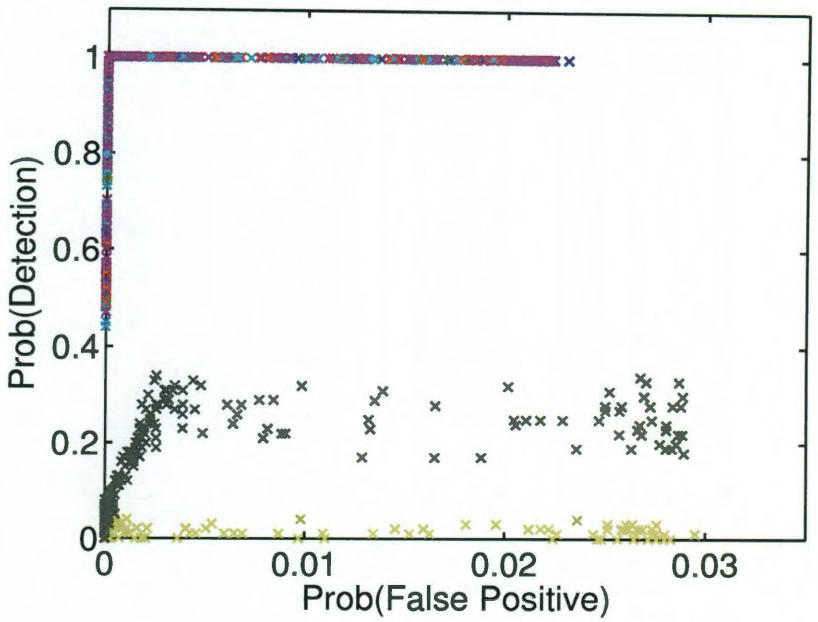


Figure 4.2 : ROC curve showing Probability(detection) vs. Probability(false positive) for random probes of length 21. The two curves in yellow and black are for the species Onions yellow and Ehrlichia ruminantium, which have small genomes and low GC content; their detection here is worse than for length 19.

which means that more energy, supplied at higher temperatures, is required to break them. With longer probes, once enough bonds are formed they stay that way instead of being in a state of flux between being bound or not. For instance, the melting temperature of a 15mer is 50°C, whereas for a 30mer it is 85°C, both with a GC content of 50-55%. However, probes that are too long are also less attractive since they are prone to sloppy binding, and their hybridization model becomes increasingly dependent on mismatch energy parameters. In summary, we want random probes to be short enough that that they stick in a sufficient number of places in each bacterial genome with a large number of bonds, and are not prone to sloppy binding, but we

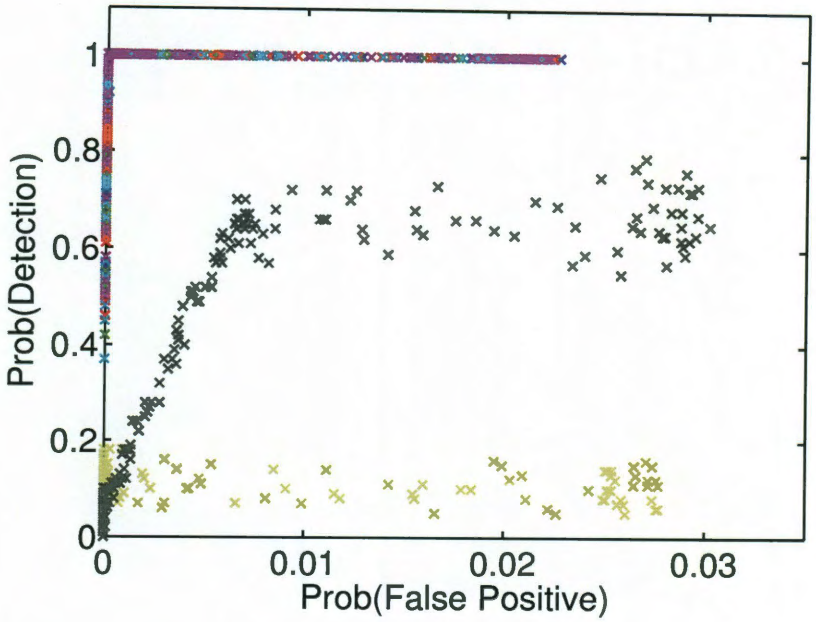


Figure 4.3 : ROC curve showing Probability(detection) vs. Probability(false positive) for random probes of length 23. The two curves in yellow and black are for the species Onions yellow and Ehrlichia ruminantium, which have small genomes and low GC content; their detection is slightly improved but shows a high false positive rate due to sloppy binding.

also want them to be long enough that they have the melting temperatures we require.

Accuracy did not change greatly even when the hybridization intensities in Φ were thresholded at values between 10% and 75%, with the underlying implication that small intensity values may be neglected in a practical scenario. From this exercise, and others similar, we conclude that as long as there are enough probes, at least on the order of the number of targets needed to be detected, there is sufficient variation in the target-probe hybridization pattern that CS reconstruction algorithms can accurately detect the target under consideration.

However, the detection power of an RPM, given the constraints in the GC-contents of probes to be used for biochemical reasons, may be restricted to genomes with either or both of a large genome and mid-to-high GC content. As an observation from simulation results, the targets with both shorter length genomes and lower GC content tended to bind with fewer random probes, rendering them more susceptible to being missed or misclassified. Our explanation for this is that in shorter genomes there is less search space for the probes to align significantly, and therefore fewer probes will be bound. We do not see the same degradation in performance for genomes that have very high GC content. There may be two reasons for this. One, in the biological world the genomes with very high GC content also often have very large genomes. Examples of such genomes in our dataset with these characteristics are Myxococcus xanthus (GC = 69%, genome length 9139763) and Bradyrhizobium (GC = 65%, genome length 8264687). Neither of them were subject to poor detection and false positive rates unlike their counterparts shown in Table 4.1. The second possible reason may have to do with the nature of the bonds themselves; we know that the G-C bond is stronger than the A-T bond, so even if probes are randomly generated and have a uniform GC distribution, their binding with a GC-rich target will be stronger and more stable than with an AT-rich one. Corroborating this idea, we noticed that GC-rich probes ($55 - 60\%$) bound to more targets on average than those with 40-45% GC-content.

Table 4.1 shows a list of the different bacteria which had lower detection rates than the others, and which also happen to be the 5 bacteria with the shortest, AT-rich genomes. The variation of their detection probability with false positive probability is show in Figures 4.1,4.2,4.3. Furthermore, of all the 100 species, the two strains of Ehrlichia ruminantium were the ones most commonly confused with one another, in

spite of the presence of several other multi-strain bacteria.

Table 4.1 : GC contents and genome lengths of bacteria with lowest detection rates

Bacterial species	GC content	Length
Ehrlichia ruminantium str. Welgevonden	27%	1512977
Ehrlichia ruminantium str. Gardel	27%	1499920
Onion yellows phytoplasma	28%	860628
Mycoplasma agalactiae	30%	877438
Borrelia	29%	910681

The promise to eliminate the need for PCR may be one of the foremost advantages of an RPM. This is highly dependent on probes binding across the entire genomes of bacteria, since only then is it possible to have enough fluorescence that there is no need to amplify any specific sections. Figure 4.4 shows an example of the spread of random probes across the genome of E. coli. The slight bias toward the sense strand in this case should not be interpreted as pathological of all probes or all targets. (The notation $5' - 3'$ and $3' - 5'$ represent the physical orientations of the *sense* and *antisense* strand of the genome respectively. The 5' or 3' refers to the termination of the strand at the 5th or 3rd carbon atom of the sugar ring.)

4.5 Experimental Design

In this section we describe how the Φ describing the affinities of random probes may be determined from experiments, including positive and negative controls. We also describe our own experimental design of a random probe-based molecular beacon

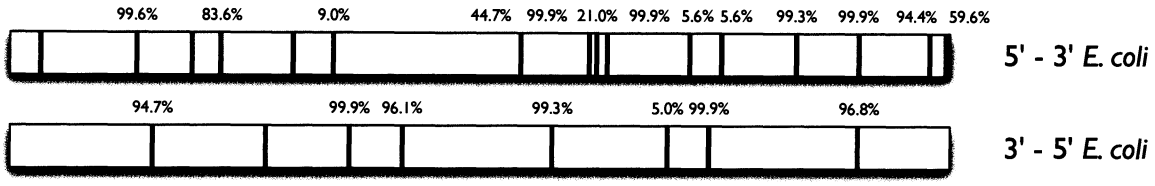


Figure 4.4 : Locations of hybridization for a sample random probe of length 19 in an E. coli genome. Percentage values were omitted in places for clarity, but the bands specify their locations.

(RMB), and the challenges yet to overcome for its validation.

4.5.1 Experimental calibration of Φ

We may desire to determine each element of Φ from experiments instead of using the thermodynamic affinity model; this is depicted in Figure 4.6. Instead of using the percent bound for each target with its corresponding probe, each element of Φ is given by the fluorescent intensity from each probe's hybridization with the corresponding target, measured in RFU's (Relative Fluorescence Units). Both probe and target concentrations remaining fixed across all Φ elements, so each Φ is defined for their fixed concentrations. With an experimentally determined Φ whose values are in fluorescence intensity, there is no step in translating from percent bound to spot intensity. During a measurement of a given sample, the fluorescent intensity of each probe is simply the weighted linear combination of the fluorescent intensities that would have occurred with each target individually. In order to estimate the true target concentration, we finally multiply the weights that the CS algorithm delivers with the original target concentrations specified for that Φ.

We also need to incorporate the positive and negative controls that are taken into

account in any set of wet lab experiments. The positive control value is determined by the intensity of the probe when it hybridizes with its perfect complement, and the negative control value is the probe intensity when it is by itself. The usual calibration method is:

$$I_{\text{calibrated}} = \frac{I_{\text{raw}} - \text{NC}}{\text{PC} - \text{NC}},$$

where $I_{\text{calibrated}}$ is the calibrated probe intensity that should be used for further interpretation, I_{raw} refers to the raw probe intensity measured directly from the experiment, PC is the measured positive control intensity and NC is the measured negative control intensity.

If this calibration is performed for every probe-target pair individually, then there will be a diminishing effect in the y signal out of the model due to too many negative controls being subtracted, as compared to the single negative control that would be subtracted in the numerator of the actual measured signal y. The difference in the model-specified Φ and experimental Φ are described in Figures 4.5 and 4.6.

4.5.2 Random Molecular Beacon Experiments

As mentioned earlier, the utilization of randomness/nonspecificity in DNA sensing need not be confined to DNA microarrays. We have also designed a set of random molecular beacons for bacterial detection. The chief advantage of using beacons instead of microarrays is that they may be easily adapted into a microfluidic device, which is more useful in a clinical setting. Microarrays on the other hand require extensive hybridization times. From the point of view of the linear model we have adopted, the physical separation of different beacons is also apt, since it eliminates any possibility of probe-probe interactions.

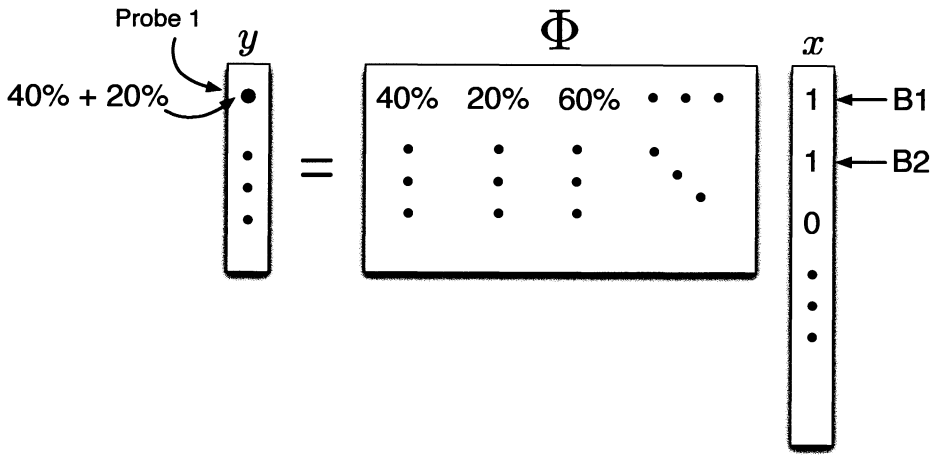


Figure 4.5 : General depiction of combinatorial probe-based sensing in the RPM, with affinities in percentages in Φ from a thermodynamic hybridization model. Each percentage refers to the affinity of a fixed molarity of a target with a fixed molarity of probe.

In each molecular beacon we designed for experimental validation, the loop of the beacon was a random sequence of length 21, while the length 5 stem sequences were consistent across all beacons. These parameters were determined after simulating across a range of lengths between 18-25, keeping the stem constant. We checked secondary structure of these beacons through the affinity model in Visual OMP DE, and chose beacons that had a stable structure where the fluorophore was certain to be quenched. To verify these beacons in simulation, we used the same two-step procedure using ThermoBlast and Visual OMP DE as when determining hybridization affinities for microarray probes.

For experimental verification, we chose a set of 3 bacteria (Escherichia coli MG1655, Francisella tularensis LVS, Staphylococcus aureus USA 300) to test our designed random molecular beacons. We ran simulations using a concentration of

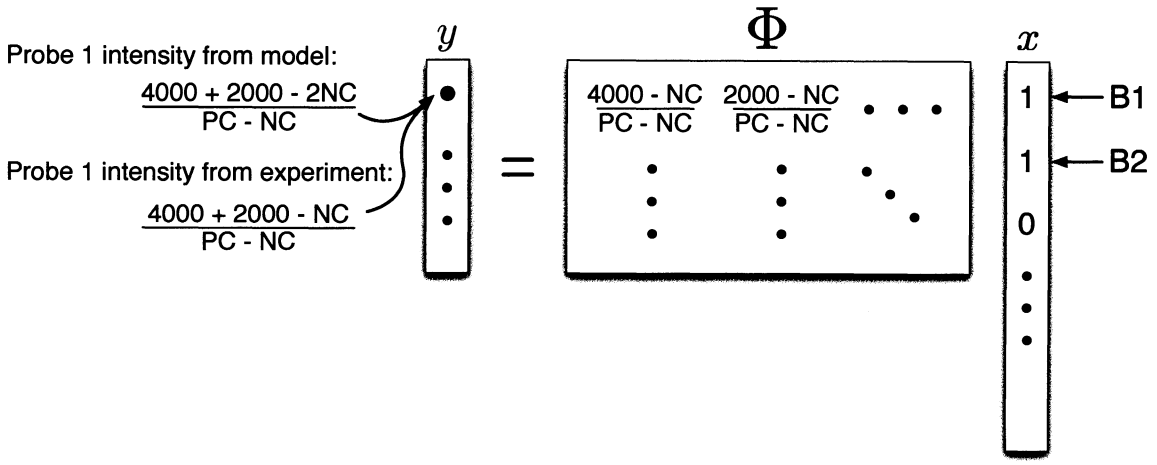


Figure 4.6 : Depiction of the experimental calibration of values of Φ from fluorescent intensities in RFU's (relative fluorescence units). There may be a slight deviation from the intensity predicted by theory compared with that in an experiment due to the negative control being added multiple times for each element of Φ.

1nM for both probes and targets, and saw that beacon fluorescence varied between 30-60%.Of these beacons, we chose the 5 with the highest fluorescence to test experimentally.

The stock DNA of the triple digested E. Coli, F. tularensis, and S. aureus we received ranged from $0.1\mu g/\mu L$ to $0.4\mu g/\mu L$. We diluted these to 0.01pM concentrations using PCR grade water. The 5 molecular beacons we used were purchased from Fisher Scientific (Thermo Fisher Scientific, Inc.), with FAM-6 fluorophores attached. These were diluted to $1\mu M$. We also purchased the complements of the loop portion of the 5 beacons to serve as an upper bound on the signal values. The beacons and targets were combined in PCR tubes and left to melt and then hybridize in a BioRad thermocycler for 5 minutes at 94°C, 3 minutes at 50°C, 1 minute at 30°C, and incubated at room temperature. Alongside the beacon-target combinations, we

also added PCR tubes for the pure beacons and the beacon-complements to serve as lower and upper calibration bounds, respectively. Finally 10μL of each mixture was placed in a 96 well plate which was excited by 485nm, and collected at 520nm in a plate reader (Horiba Scientific fluorolog). The resultant intensity values were analyzed. Besides the concentrations indicated above, this procedure was repeated for other varying beacon and target concentrations, and thermocycling conditions.

Unusually, the beacons we tested showed signal values that were below the noise floor. However, when $MgCl_2$ was added before the thermocycling process, instead of before the excitation process, all signal levels dropped – especially the noise floor which dropped drastically. This indicated to us that the beacons were unstable even in isolation, possibly due to weak binding in the stem that should have kept the fluorophore quenched. We also observed that the weak beacon signals did not show strong correlation with the predicted values. Our explanation for this is that either: (1) the beacon signal is too weak to be seen when bound at such low target concentrations, (2) the beacon signals were obscured by the rest of the target genome fragments, or (3) the experimental protocol was such that beacon-target binding broke off and/or the target fragments bound to their complements instead.The plate reader itself is sensitive to fluorescence from even 1-2 molecules, so we believe it is not the equipment at fault. We tested this hypothesis by immersing a perfectly matched, fluorescent, FAM beacon-complement pair in diluted E. coli genome, which extinguished the fluorescence. Therefore, one direct implication of this set of experiments is that there is a definite need to increase the signal strength of the beacon.

One simulation-based change that we attempted was to find beacons with different structures to ensure greater coverage in the target genome. Every target genome is fragmented on the order of 30,000 fragments, of which only $\sim$1000 fragments have

relatively strong hybridization with the molecular beacons. We recognize that the key to stronger fluorescence is not finding beacons that increase hybridization intensities in any given beacon-target fragment hybrid they form, but rather, creating *more* beacon-target hybrids. For this reason beacons with shorter loop and longer stems may be more appropriate, for greater target coverage from the short loop and increased beacon specificity in binding from the long stem – while also decreasing the noise floor by creating more stable beacons. However, from simulations we saw that our hypothesis was incorrect thermodynamically, since the number of bonds formed between the shorter loop and each part of the target genome where it bound were not sufficiently high energy to break the stem bonds. As a result, beacon binding was poorer than in the previous case. We are also restricted to using beacon loops of 18-25 base pairs so that their melting temperatures are neither too high nor too low.

In our second set of experiments we ordered new molecular beacons where, instead of a quencher, there are two fluorophores Cy3 and Cy5 at either end of the stem. Cy3 and Cy5 have stronger fluorescent intensities than FAM fluorophores. When the beacon is in its default state it emits mainly through Cy5 (at 668nm) through FRET (Fluorescent Resonant Energy Transfer) from close proximity to Cy3 when excited (at 485nm). When it is in a bound state the molecular beacon opens from target binding, and it emits only at Cy3 (564nm) due to the increased distance between Cy3 and Cy5. The E. coli target that we used was fragmented using a hydroshear machine ensuring more uniform fragmentation into fragments of 300-400bp compared with the previous set of experiments that used restriction enzymes for fragmentation. The same experimental protocol was carried out, but modified in the annealing step, so that the samples were left to cool at 30°C for 1 hour, allowing sufficient time for any beacon-target hybridizations to take place. The positive control and E. coli

target were both tested at a variety of concentrations ranging from 10^{-9}M to 10^{-13}M. Fluorescence was read using the same plate reader as before.

The results of the second set of experiments showed good SNR for the positive control up to a concentration of 10^{-11}M, but little to no signal change for the E. coli target. Revising the results from our first set of experiments, this tells us that the beacon signal should be strong enough to bind at low target concentrations, and the fluorescent intensity sensitivity of the plate reader is adequate. The remaining challenges to our experiments, assuming our theoretical affinity model is accurate, are: (1) the beacon signal is obscured by the rest of the target genome fragments or (2) the experimental protocol is not appropriate, so all beacon-target bindings broke apart, and the target fragments bind to their complements instead.

The solutions to these are as follows:

- to anchor the beacons to the slide and wash away the rest of the target fragments that may be obscuring the signal. This may be done by coating the slide with streptavidin, and adding a biotin linker to the molecular beacon. Then, after hybridization excessive target molecules can be washed away leaving only the fluorescent beacons.

- to change the experimental protocol to allow still more time for annealing of beacon-targets instead of target duplexes, and

- as a last resort, to generate new beacons with still higher melting temperatures, if possible at all

One immediate change to make is to introduce a PCR step amplifying only certain sections of every bacterial genome using shared sites for binding of the primers, and

to simulate and generate random probes for those regions alone. Then the fluorescence signal intensity will be much stronger due to many more beacon-target hybrids formed. Even if some beacon-target hybrids break up during the protocol, there will be such a large number of them formed in the first place, so that the signal is still visible. While this will indeed validate the use of random probes, it will also obviate one important advantage of using random probes that we set out to accomplish in the first place – obviating the need for PCR, which would add to the time and cost of such an experimental protocol in a clinical setting. However, we would still be able to identify as many targets with as few beacons, and the same random beacons would also have affinity for unknown targets (assuming they also possessed the same PCR primer binding sites), fulfilling our goal of nonspecific sensing.

4.5.3 Extensions to other molecular devices

Finally, we mention two other potential extensions to existing molecular devices that may be adapted for bacterial identification. The hybridization prediction affinity model (Φ) that can be used is the same as for the RPM or Random Molecular Beacon.

- Random NanoStrings: NanoStrings® (NanoString Technologies) are a novel method for direct multiplexed measurement of gene expression that use molecular bar codes followed by single molecule imaging to detect hundreds of unique transcripts in a single reaction. Each target molecule is identified by two NanoStrings; these are unique probe sequences, one of which is attached to a linker molecule, and the other, to a color barcode. After hybridization the probe-target-probe hybrids are immobilized via the linkers, the remaining DNA is washed away, and the bar codes are imaged and counted. In this way, it is possible to literally count the number of times a certain DNA sequence may

occur in a target. We propose the use of random NanoStrings: non-unique, probe identifiers, attached to barcodes, which together create a combinatorial identification system for each gene.

- Nonspecific primers in qPCR: While qPCR is a precise and sensitive technique for DNA detection and quantification, it is expensive, and organism-specific. Testing for a certain pathogen requires a specific kit/reagents to be purchased for that pathogen. We propose the use of random, nonspecific primers instead, which may be able to hybridize with and amplify many different pathogens. As with the RPM, a group of such primers will characteristically amplify each pathogen target, which is detected and identified in real-time. One design we envision is a "random" Scorpion™ structure, where a molecular beacon is attached to a primer; one or both sequences may be randomly generated, and together they form an appropriate length for creating characteristic bacterial signatures. The same set of random Scorpions may be used to detect any pathogen.

4.6 Real world considerations

4.6.1 Application scenarios

It may sometimes be useful to design a random probe platform specifically for an application scenario, instead of using a one-size-fits-all platform. We categorize its uses into single and multiple bacteria detection scenarios.

- Single bacterium diagnostic: used for bacteremia, biodefense

  In the normally sterile blood, the occurrence of a single bacterial species (medically known as bacteremia) is typical. Similarly, in the case of engineered or

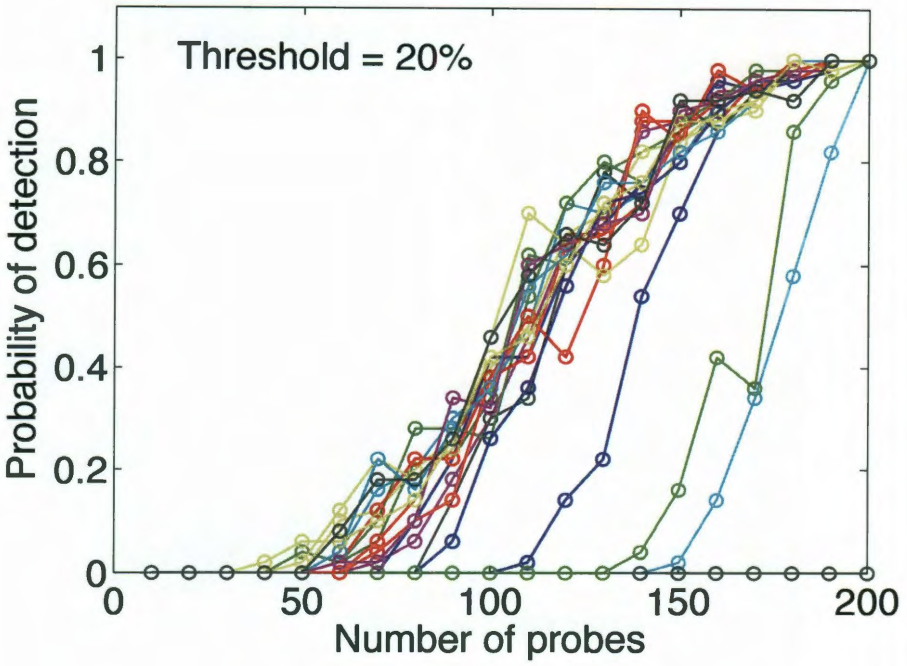


Figure 4.7 : Probability of species detection varying with the number of probes (M) using a thresholded at 20%, binary Φ as our perturbed Φ for decoding. This plot shows the detection curves for 20 randomly chosen bacteria out of $N = 100$. Notice that the only species that cannot be detected at $M = 200$ in this set of 20 is E. ruminantium with GC content of 27%.

naturally occurring pathogens used in bioterrorism attacks we may expect only a single species in a sample (such as an envelope of anthrax). In the case of single bacterium detection we can minimize the number of probes further than for other scenarios that require multiple bacteria. The target set and example simulations that we have followed in this thesis showed that we could detect 100 bacteria (excluding 6 AT-rich bacteria) with as few as 50 probes.

This type of random probe platform that does not require PCR and supplies

fluorescence intensity results within a few hours (as with molecular beacons) would be useful to clinicians for rapid diagnosis of the infection without culturing – regardless of its ability to detect unknown species. Species information will also inform the antibiotic susceptibility testing procedures to determine which antibiotics those bacteria will respond to. A similar platform may be built for viruses, which have smaller genomes than bacteria.

- Multiple species analysis: human microbiota, environmental samples

  At several sites where microbiota reside on a healthy human body, such as the oral cavity, gut, or even the epidermis, they reside in the form of multiple, diverse, species. The gut may contain on the order of 100 different bacterial species. Similarly, air, water and soil environments that constitute the biosphere are rife with bacterial variety. Understanding the species populations in these complex environments (known as the *metagenome*) is a research task, whose insights we can derive health and pharmacological benefits from, rather than use in a clinical diagnostic.

  Sequencing is a popular tool in such situations, but is prone to high error rates. Here, a large random probe platform may be useful since it can identify the occurrence of multiple species in a sample, without potential nonlinear amplification due to PCR. The larger the random probe set, the more accurate the identification of both known and unknown bacterial species will be. On a random probe platform, since the probe concentration is in far excess of the target/species concentrations, we are able to linearly combine the probe signatures that were measured for individual targets using the hybridization model. Therefore even if multiple species are present, the random probe platform will still

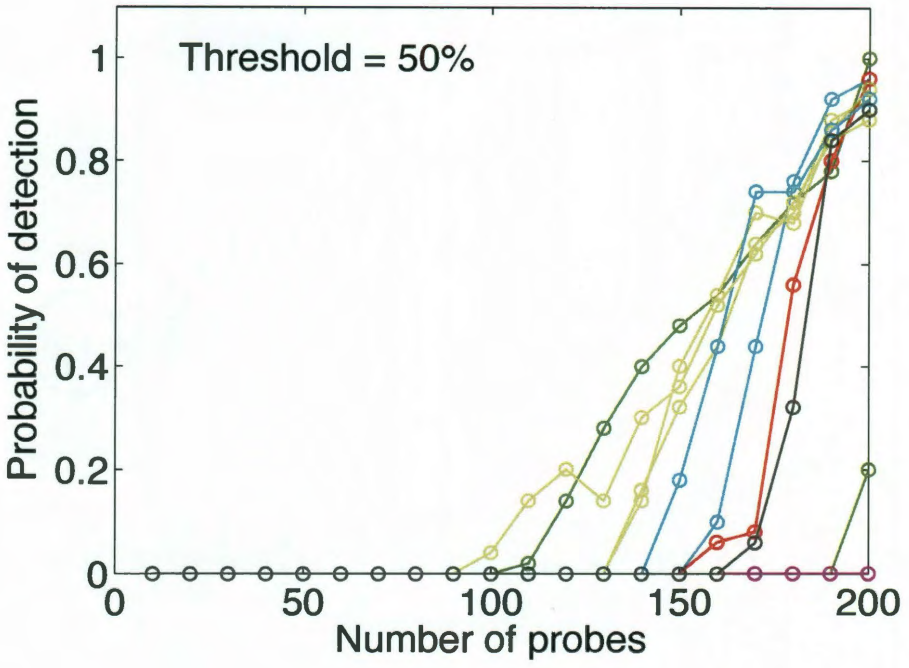


Figure 4.8 : Here Φ values are thresholded at 50% and then converted to binary. We see decreased detection performance due to greater perturbation; now only 8 of the 20 bacteria can be detected with the same set of probes. ϵ_Φ here is 0.82, compared to 0.40 in Figure 4.7.

perform accurately but may need a larger number of probes as shown through our simulation results. In any sample for metagenomic analysis, species concentrations are at the levels (lower than 10^{-10}M) appropriate for our linear model.

4.6.2 Perturbations to Φ

We have taken great care to use physically accurate models that reflect the spot intensity from each probe-target pair hybridization. We have also described the exper-

imental control-based calibration of each probe-target pair for even greater accuracy in spot intensity values. However, in spite of these precautions, errors due to our model may enter into our prediction for Φ. We can evaluate the effect of these errors as perturbations to Φ. As described in [30], we characterize this as some perturbation matrix E added to Φ, so that $\hat{\Phi} = \Phi + E$. For our situation, E is due to inaccuracies in our estimation of the spot intensities in Φ, resulting in the model-based $\hat{\Phi}$ that we use for decoding.

However, experimentally, probe-target hybridizations will occur as prescribed by their true affinities in Φ, and will determine probe measurements in y. As shown in [30], the recovery error when observations are taken using Φ, but decoded using $\Phi + E$ using a sparsity-based method (Basis Pursuit), will scale linearly with the magnitude of the relative perturbation ϵ_Φ, where $\|E\|_2 = \epsilon_\Phi . \|\Phi\|_2$.

It is difficult to estimate the magnitude of the perturbations in our model-based $\hat{\Phi}$ *a priori*. Instead, here we choose a perturbation that may be adopted whenever we are unsure of our model: threshold low intensities of real-valued Φ and convert the matrix to be binary-valued. This matrix then becomes the $\hat{\Phi}$ that is used in decoding, and is a crude model for spot intensity prediction. To evaluate the robustness of this thresholded-binary perturbation through simulation, we allow our measurements to be formed by the original real-valued Φ, and decode from them. In general, any adverse effects from greater relative perturbations ϵ_Φ on detection are countered by the large number of random probes that a random probe platform uses compared to the number of targets. Recall that in all figures previously described in this chapter, 200 probes were used to detect 100 randomly chosen targets with a binary valued Φ that, unlike here, was used *both* to create measurements y and for decoding. Mathematically, we can show our use of perturbed Φ as:

$$\hat{\Phi} = I_{\geq \text{thresh}}(\Phi) \tag{4.1}$$

$$y = \Phi x \tag{4.2}$$

Here, $I_{\geq \text{thresh}}$ is an indicator function for all values above the threshold *thresh.* y is created using the true affinity model Φ. We solve for the solution x^* using y and $\hat{\Phi}$:

$$x^* = \min_{\hat{x}} \|\hat{x}\|_1 \quad \text{s.t. } y = \hat{\Phi}\hat{x} \tag{4.3}$$

for some $\epsilon > 0$.

Numerical simulations using this thresholded binary perturbation model are illustrated in Figure 4.7. This plot shows the detection curves for 20 randomly chosen bacteria out of the same set of $N = 100$, and how their individual detection varies with the number of probes. Here, the Φ used in decoding is binary-valued and thresholded at 20% intensity. As we would expect, with increasing number of probes, M, we are able to better detect each species. The only species that cannot be detected at $M = 200$ in this particular set of 20 is the AT-rich E. ruminantium. This plot tells us that in spite of our crude perturbations to the true affinity model we are still able to achieve good detection performance at the expense of a larger number of probes. ϵ_Φ here is calculated to be 0.40, implying that the mean squared error in sparse recovery using $\hat{\Phi}_{20\%}$ is 40 times larger than that if we had used a $\hat{\Phi}$ with ϵ_Φ equal to 0.01.

We can contrast this for a situation with greater perturbation; suppose we use a threshold of 50% instead. In this situation, as illustrated in Figure 4.8, detection performance decreases greatly. We calculate ϵ_Φ in this case to be 0.82, implying that we will see MSE of x^* increase by a factor of 2 compared to the case with a 20% threshold. However, for the purposes of bacterial detection, we are interested in

detection probability rather than MSE, which we plot and see that it is also adversely affected.

Chapter 5

Sparsity-based Methods for Bacterial Sequencing Data

5.1 Motivation

The number of bacterial cells in and on a human body outnumber the actual human cells by at least a factor of 10. Understanding the composition of the microbiome in healthy adults, and contrasting it with its other conditions may help us develop new prognostics, diagnoses and cures. For instance, Helicobacter pylori-induced gastritis is known to be the strongest singular risk factor for cancers of the stomach, but only a few strains exhibit the proteins that cause malignancy [31]. In other situations, it is useful to analyze environmental samples to detail the habitats of certain species. Such deep investigations studying *metagenomes* require more precision and detection capabilities than hybridization-based microarrays or other coarser hybridization-based tools may provide, so they instead undergo the gold standard for genomic assessment – sequencing.

Even after sequencing, the identification of bacteria from sequencing data still depends on trawling through it for the 16S, 23S or other genes that have enough variation between species so that they may be uniquely mapped. Often the experimental sequencing process is modified to include the isolation of such identifier genes and their PCR amplification, followed by the actual sequencing process and subsequent data analysis. Put together, all these steps create an arduous processing pipeline.

Recently, scientists have turned to the concept of Whole Genome Sequencing (WGS), where 16S or 23S gene identifiers are not isolated from a sample; instead detection is performed on complete bacterial genomes. (In the case of environmental samples, this is called environmental shotgun sequencing.) This type of direct genomic assessment usually precludes the need for PCR, or as many cycles of PCR, since there is no single gene identifier we are targeting. However, the time saved in skipping experimental steps has been offloaded to the data analysis side which must now draw correlations between DNA fragments across each genome instead of specific to a gene. For example, generating 100 million 90mer long reads can take 8-9 hours on an Illumina GA_{II} sequencer, while running those millions of reads against the thousands of sequenced bacterial genomes in the NIH database, even using state-of-the-art algorithms (BLAST – Basic Local Alignment Search Tool) can take upwards of 24 hours running on 20 processors. It is therefore imperative to devise faster identification algorithms for WGS, especially for multiple species detection. As real-time sequencing instruments become cheaper and more ubiquitous, the need for faster sequencing data analysis will only grow, and real-time algorithms to complement them will be more critically required.

5.2 Proposed solution

Our main goal in sequencing data analysis is an improvement in speed, without sacrificing accuracy. Therefore it is beneficial to develop a solution that enables as much preprocessing as possible before parsing raw sequencing reads against it. Our proposal is to model every sequenced bacterial genome in the NCBI database by the frequencies of occurrence of the *kmers* in them, and compare it with the kmer frequency distribution in sequencing data to decipher which bacteria occur.

The kmer-frequencies in the sequencing reads from multiple bacteria in a sample are linearly modeled by the kmer frequencies in each bacteria, weighted by the number of bacteria in the sample. The proposed method allows a good deal of preprocessing in the development of kmer-based genome models, and can work from raw sequencing reads, instead of an assembled contig from a genome alignment and assembly step first. (It is also excepted from the vagaries of these prior algorithms in the processing pipeline.)

We observe that the number of bacterial species that may occur in a given sample is small compared to the total number of sequenced genomes in our database, so the problem is a sparse detection problem. The use of sparsity guides us to the use of a sparsity-based reconstruction algorithm. By the virtues of Compressed Sensing theory, using a sparsity-based model will mean that we need fewer sequencing reads to generate an accurate solution. Even though data minimization is not a primary objective of this analysis, it is an important general consideration for all sequencing technologies. Furthermore, fewer data to analyze also directly translates to a gain in analysis speed.

Binary representation

We convert all our sequencing data and genomes to a binary representation, by representing a single nucleotide as a "1" and all others as "0"'s. We refer to this as its nucleotide skeleton, as per which that nucleotide takes the "1" value. For instance, the T-skeleton of a sequence reading "ATTCGT" would be "011001". Our intuition is that there is enough diversity in the distribution of even a single nucleotide in a genome that it can be used to distinguish between different genomes.

This modification will exponentially reduce the complexity of our model (and

consequently the amount of data it would require). In binary the number of possible length 10 kmers, or 10-mers is only 2^{10} which is 1024, whereas in its original notation, there would be 4^{10}, approximately 10^6 possible. The number of kmers occurring in the genome plays a significant role in our model, so it is important to limit where possible. Moreover, we see that an A-C-T-G-based model falls prey to the same limitations that our binary model does. In the analysis that follows we primarily use the T-skeletons of genome.

Note that the correlation structure in a nucleotide-skeleton of a double-stranded genome is identical to that of its complementary nucleotide; i.e. we do not gain any more information from analyzing a T-skeleton than we do an A-skeleton, for example. At most, we may choose to use information combined from a T-skeleton and a C (or G) skeleton. But the corresponding increase in the kmer frequencies due to this is only a multiplicative factor of 2, whereas using A-C-T-G representation would mean an increase by an exponential factor of 2.

5.3 Previous work

There has been significant work in the last few years on the problem of metagenomic analysis for environmental samples. The methods closest to ours use PCA (Principal Component Analysis) based linear dimensionality reduction by feature selection, followed by a linear classifier such as LDA (Linear Discriminant Analysis), to categorize each kmer-binned sequencing read according to the bacterium it came from [32, 33]. Both approaches use hexamer-binning (kmers of length 6) of reads of length 1000. The chief limitations of the approach in [32] is that it cannot actually identify any bacteria itself, but instead projects the sequencing read data into a lower dimensional feature space, and then uses those same new features to categorize each read as being

from one bacterium or another. With this approach the authors are able to separate between 2-6 bacteria using synthetically generated sequencing reads from their genomes. The approach in [33] can identify the genomes associated with reads, but requires a training phase for its classifier, making its performance dependent on its selection of training data and how much it is allowed to train for. Furthermore, both algorithms make use of little preprocessing, resulting in longer online computational times.

Kmer-frequencies are in fact not a new idea in species detection and have previously been used to try to identify unique genomic signatures. But many kmer-frequency binning methods suffer from two limitations, as noted in [32]. One, they have been known to perform poorly on shorter sequencing reads, and are consequently applied to assembled contigs. Two, they require an alignment or training phase against current genomes. Our suggested method works the same irrespective of sequencing read length, and does not require any alignment of reads against reference genomes. The kmer-genome model only needs to be determined once for a given set of bacteria and kmer length, so does not contribute to processing time of the data.

5.4 Linear Sparse Approximation Setup

The WGS analysis problem involves parsing a list of $\sim$100 million reads to decipher which bacteria are present in a sample. We may consider each bacterium to be a linear combination of kmers; the kmers corresponding to a bacterium are determined by passing both the sense and antisense strands of its genome through a k base-long sliding window and counting how often they occur; then each kmer's weight in the genome is exactly its frequency of occurrence in it. The sequencing data that we analyze is a linear combination of different bacterial genomes, weighted according to

the number of copies of each genome present in the sample – and therefore itself also a linear combination of kmers. The relationship between kmers in the sequencing data and each bacterial genome is depicted in the following linear system.

$$y = \Phi x \tag{5.1}$$

Here, Φ is an $M \times N$ matrix essentially composed of histograms of the M unique kmers that occur in a set of N bacteria. Each element $\Phi_{i,j}$ indicates how many times the ith kmer occurs in the jth bacterial genome. The y vector is of length M, obtained by tallying the frequency of each of the M kmers in the millions of sequencing reads. Therefore, there is no dependence on actual lengths of sequencing reads (which is determined by the number of sequencing cycles that are specified during the sequencing process). Read lengths are typically 45-90 bases long, so even a few thousand of them produce enough data to parse a large number of 25mers from it. The reason for generating such volumes of reads is to have enough coverage and redundancy in covering the entire genome multiple times. Even the largest bacterial genomes are on the order of 10^7 bases. In equation 5.1, each element of y corresponds to how many times that kmer occurred in the reads list. The objective in the system is to solve for x, which is of length N; it indicates how many copies of each bacterium were present in the sample. x, y, Φ are all nonnegative integer valued.

It is safely assumed that x is sparse – few bacterial species are present in any given sample relative to the number of sequenced bacterial genomes known, so that $K \ll N$. The problem then becomes a sparse approximation problem, similar to Compressed Sensing except that the system here is considerably overdetermined; typically M can be $\sim 10^6 \times$ N. Current estimates for the number of sequenced bacterial genomes put it at < 3000.

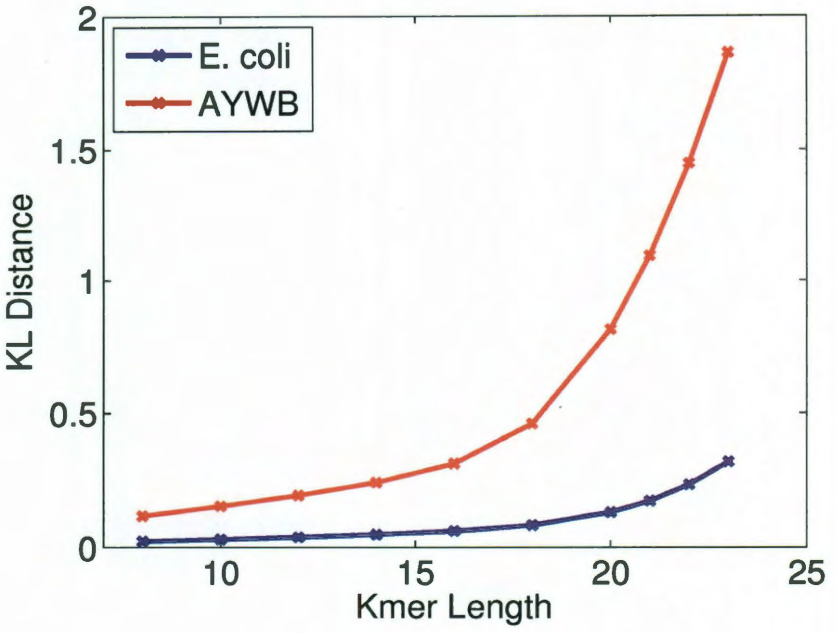
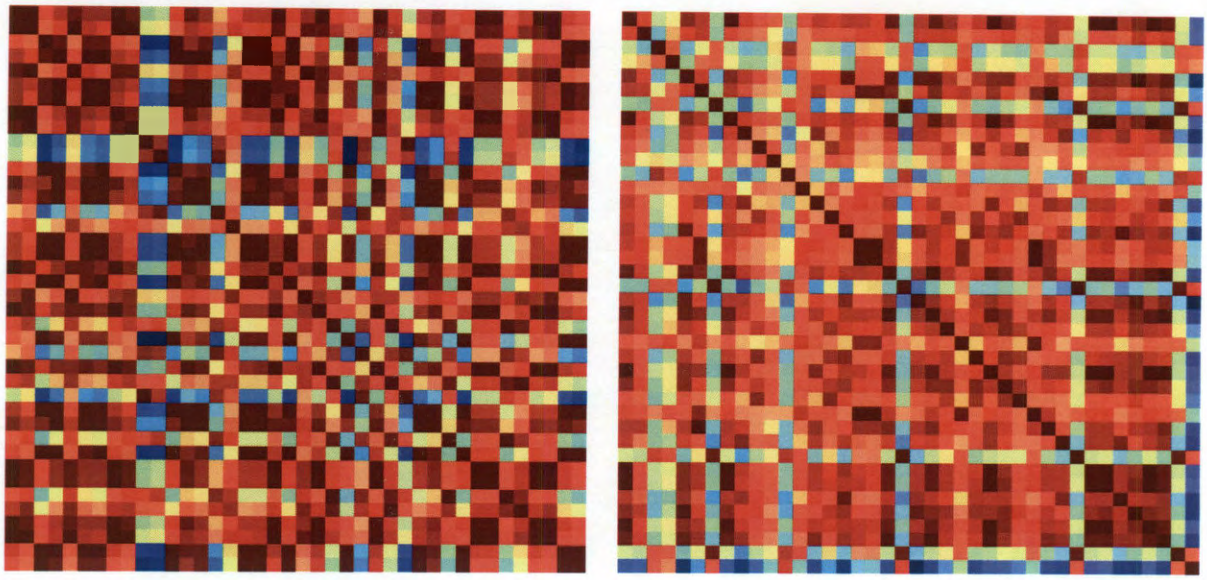


Figure 5.1 : Variation of the KL distance between the kmer frequency estimates of the genomes of E. coli and Aster Yellow Witches' Broom (AYWB) from their random iid counterparts. AYWB has a much smaller genome generating fewer samples than E. coli for a given kmer length, and diverges faster and greater from its iid distribution.

5.5 Kmer Frequencies in Genomes

The choice of kmer length plays an important part in this model, as the kmer-based frequency distribution of each genome corresponds to its column in Φ, and we prefer them to not be highly correlated with one another so that our sparsity-based reconstruction algorithms can be applied with guarantees. We observe that when k is too short, the probability mass function estimates of kmer occurrence in a genome tend to look similar to those of its random iid counterpart, given by the same T-content distribution and genome length as the real bacterial genome. We also observe that



(a) (b)

Figure 5.2 : (a) Grammian of the Φ based on 10mer estimates for a set of 40 randomly chosen bacteria (b) Grammian of the Φ based on 10mer estimates with its iid 10mer estimates subtracted.

given any kmer length, longer genomes tend to look more similar to their iid kmer distributions than shorter genomes. Figure 5.1 shows the increasing KL distance* for two different bacterial genomes with increasing kmer length. The genome that is much shorter, AYWB, of length 706569, diverges from its iid distribution much faster and in larger amounts than the much longer E. coli genome, of length 4639675. This effect has to do with the number of samples used to fill in a given kmer distribution for a genome – shorter kmer lengths and longer genomes mean more samples in its

*The KL (Kullback-Leibler) distance is a measure of the distribution between two probability distributions, P and Q, where Q typically represents the theory or approximation of P. It is calculated as $D_{KL}(P\|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$.

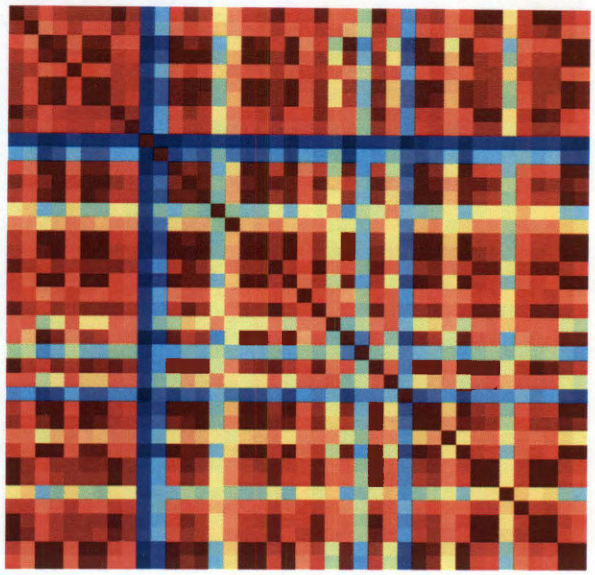
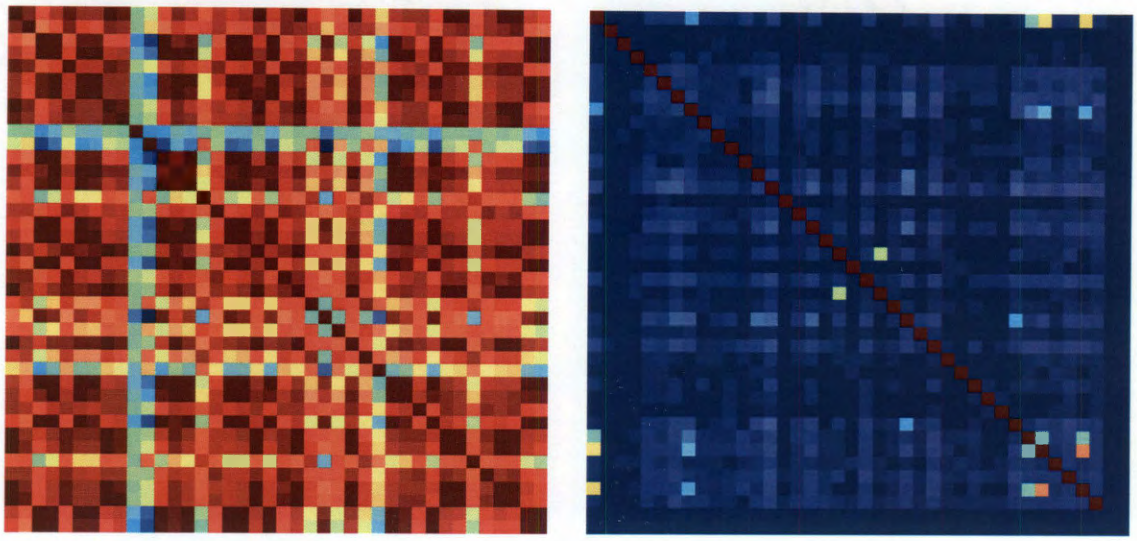


Figure 5.3 : Grammian of the Φ based on 25mer estimates for the same set of 40 bacteria.

kmer distribution. The shorter the kmer, the more sample points can be obtained from it for any given genome. Similarly, when the genomes are longer, they generate more kmers for a given kmer length.

As it turns out, the more samples used to describe a density estimate (in this case, kmer probability mass function (PMF) or the kmer frequency estimates divided by the number of kmers) of a sequence (in this case, the genome) the more it looks like the density estimate of its iid counterpart (in this case, iid nucleotides with probabilities of occurrence equal to the nucleotide-contents in the genome). More formally, the probability density estimate of a strictly stationary, correlated sequence asymptotically has the same distribution as the density estimate when sampling from independent random variables [34]. Here, the independent random variables (1's, 0's) are from an iid distribution with the same probabilities of occurrence (Prob(T),



(a) (b)

Figure 5.4 : (a) Grammian of the first 10,000 rows with lowest T content of the 25mer Φ (b) Grammian of the last 10,000 rows with highest T content of the 25mer Φ

Prob(A) + Prob(C) + Prob(G) respectively) as in the true genome. Asymptotically as the number of samples (in this case the number of kmers generated for its kmer PMF) grows, their distribution estimate looks like that in the independent case. By drawing intuition from this theoretical result, we are implicitly assuming a strict stationarity in the distribution of variables, which is not true in genomes, but allows us to use a simplified iid distribution for each genome where the independent variables all have the same T-content distribution. In real genomes there is no stationarity, and correlation structure may change depending on region; there may be some regions which are characteristically T-rich, and others that are not. However, empirically we see that this assumption still holds merit, from trends like those in Figure 5.1. Stronger correlations in the genome that are not stationary actually work to our benefit, since they imply a slower convergence to the independent case, retaining

their unique correlation structures in their kmer frequency distributions, and allowing greater coherence between columns of the Φ they generate.

From an analysis point of view, a popular heuristic to check the appropriateness of a Φ for sparsity-based recovery methods is to check the structure of its Grammian $- \Phi' * \Phi$, after column-normalizing Φ. This is an indicator of the RIP, whose verification is a combinatorial problem over every column subset in the matrix, but is a sufficient condition to guarantee exact sparsity-based recovery for a certain number of measurements (in our case, these are kmer frequency estimates). The values of the Grammian indicate the similarities between pairs of corresponding columns in the Φ, in our case a measure of the similarity between the kmer distributions of bacterial genomes. A "good" Φ is one that has strong diagonal structure, implying that every bacterial column is highly correlated with itself, and much more so than its correlation with any other bacteria in a set. If a bacterium shares high correlations with other columns, they may be confused with one another during sparsity-based recovery.

Figure 5.2(a) shows the Grammian of E. coli for kmer length 10, in a set of 40 randomly chosen bacterial genomes. We see that there is hardly any diagonal structure, and in fact all columns are highly correlated. This is due to the large sample number effect, where all genomes begin to have distributions that converge to those of their iid counterparts, as described above. For instance, we see in row 1/column 1, which is E. coli (T content = 24.6%), it is highly correlated with elements 2 and 8, which are Archaeoglobus fulgidus and Shigella flexneri, with T-contents of 25.6% and 24.5% respectively. They all have the same iid counterpart (given by the same T-content) so their kmer frequency distribution estimates are very similar to one another. Furthermore, when we subtract the iid 10mer distributions of each of the genomes from the original Φ, the diagonal structure is more strongly apparent,

as shown in Figure 5.2(b).

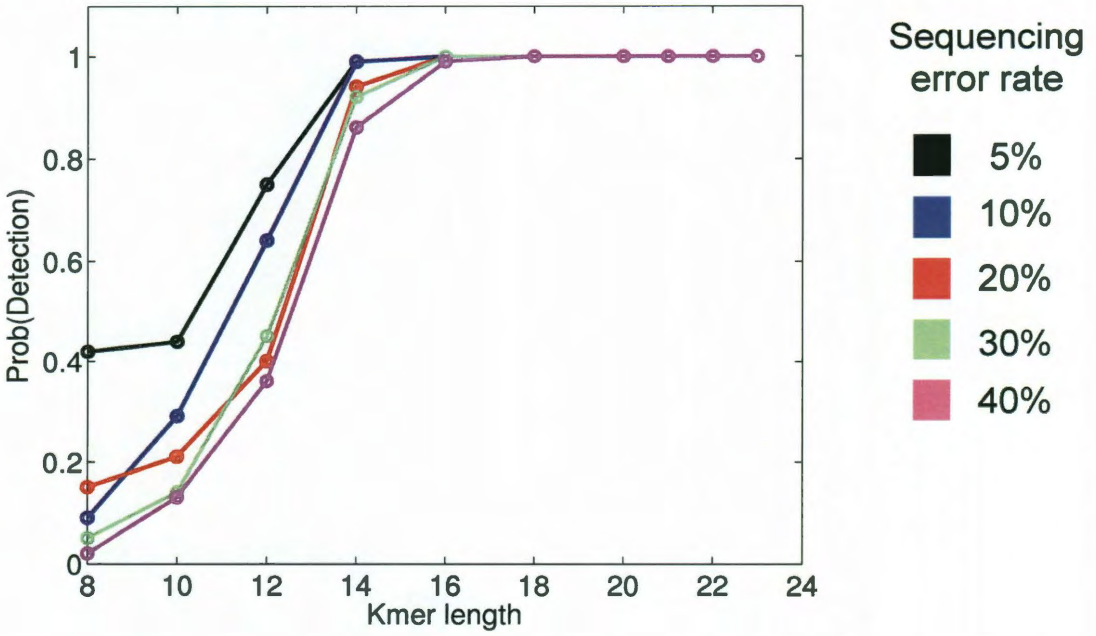


Figure 5.5 : Detection probability increases with kmer length, for changing sequencing error rates of 5, 10, 20, 30 and 40%

As kmer length increases, the Grammian structures improve; the Grammian for the same set of 40 genomes at kmer length = 25 in Figure 5.4(a) shows slightly better diagonal structure than that we saw for length 10. This improvement is due to the varied structure of some rows in its corresponding Φ which now has approximately 15 million rows compared to the 1024 in the case of kmer length = 10. For any genome's kmer distribution, the number of kmers that populate it is given by counting the kmers from the sequential sliding window on the genome and equals the genome length - kmer length + 1, which is approximately equal to the genome length in the case of small enough k relative to it. Therefore with a binning of 15 million kmers vs. 1024, we would expect there is a much wider spread for the sample points to occupy

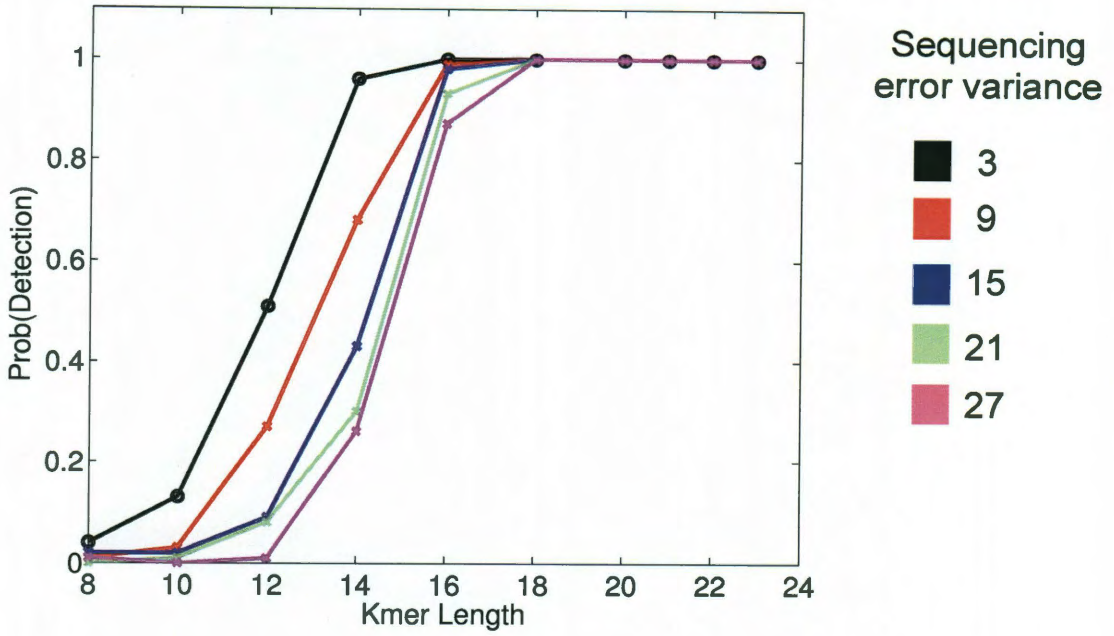


Figure 5.6 : Detection probability increases with kmer length, for changing error magnitude variances of 3, 9, 15, 21 and 27. Variance of 3 corresponds to an SNR of 2dB.

on average, resulting in many unoccupied bins. However, there is a probability bias toward kmers that have low T's by virtue of the T-skeleton distribution, and therefore many of the sparsely populated bins occur in the kmer regions of high T-content. Across a set of genomes therefore, these sparsely-occupied kmer bins are typically the ones that are unique identifiers for the genomes – they are only nonzero-valued for the kmer in which they occur. From a Compressed Sensing/sparsity perspective, their measurements are differential in their sensing of the genomes, but not holistic in that they only occur in a single bacterium. Therefore, errors in unique identifiers are not robust. For compressed sensing to occur with fewer measurements than targets, this is not an option. In our case, minimizing number of measurements is a goal,

but it is secondary to accurate and fast detection, so in that sense unique identifiers are acceptable. However, while these kmer measurements are surefire indicators of a bacterium, from a practical sequencing perspective it also means that their actual measurements in sequencing data are far less frequent, and if there are errors in their measurements, may contribute to the implication of the wrong bacterium in the sample. It is for this reason that detection methods, irrespective of the assumption of sparsity, that *only* use unique indicator kmer bins are prone to error. The 16S, 23S ribosomal DNA unique identifiers that are typically used in detection methods are prone to these same errors.

It appears that the Grammian diagonal structure is so poor that we may need to go to very large k to achieve a suitable measurement matrix. However, now we make the observation that column incoherence is not uniform across all rows due to the nonuniform distribution of binary-valued kmers. We can leverage the incoherence in different rows of Φ without needing too large a k. Figures 5.4(b) and (c) show the structure of the Grammians using only a subset of 10,000 rows, taken from the lowest and highest T-content regions respectively of each genome's 25mer frequency distribution, which are also the highest and lowest probability areas of the distribution respectively. We see that the diagonal structure using the high probability 25mer bins is poor, and very similar to that in Figure 5.4(a) where the Grammian is taken using all 15 million rows. This is because these 25mers occur with high probability in *all* genomes. However, the diagonal structure using rows in the tail of the 25mer distribution is excellent, since this is where the unique indicator kmer bins reside. In fact, the last 2 bacterial columns using this bottom 10,000 row subset are all zero, which means that none of those 10,000 kmers occur in either of their genomes. If a recovery algorithm were run using such exactly this subset, neither of these bacteria

would ever be detected. We will take care to check for this when sampling the rows for kmer detection in sparsity-based recovery methods used below.

5.6 Using Kmer-based Φ's for Detection

We describe trends in the detection capabilities of the various kmer-based Φ's in the detection of E. coli str. MG1655 from a set of 40 randomly chosen bacteria, including three other E. coli strains, and other closely related species such as Shigella flexneri. We used the popular sparsity-based algorithm, CoSamP, but others that will yield similar results include IHT and LARS with LASSO constraints [8, 35, 36]. However, instead of using the full Φ's for each kmer length, we use a biased sampling of the rows in each Φ. Each iteration of the detection algorithm used a subset of only 5% of the rows of the Φ, with 80% of them from 10% of the rows with the highest T-content in the Φ.

We introduce two main types of errors (besides the addition of low variance Gaussian noise) that we believe encompass the major error possibilities in sequencing. One, we use a sequencing error rate, which specifies the percentage of bases that are sequenced incorrectly. For many sequencing instruments this can vary between 1% and 5%. In our case, we interpret this as follows. If a single base is misread during sequencing, all kmers that contain it are also erroneous. In our parsing of sequencing data, we only take *one kmer per read.* Therefore the percentage of wrong bases in sequencing data is translated to a percentage of wrong kmer estimates in our frequency distribution. Figure 5.5 shows the variation of detection probability using different kmer lengths (between 8-23) and different sequencing error rates of 5, 10, 20, 30 and 40%).

Two, we specify a sequencing error variance, which controls the variance of the

values assigned to the errors affecting the kmer bins, which are in turn specified to be erroneous by the sequencing error rate fixed above. These values are sampled from a normal distribution and scaled by the specified sequencing error variance as well as the average energy-per-bin of the kmer estimates of the sequencing data. This scaling can be interpreted as an SNR (signal-to-noise-ratio)[†] imposition, since our noise variance/sample scales with the energy/sample. For instance, a sequencing error variance of 3 corresponds to an SNR of approximately 2dB. Figure 5.6 shows the variation of detection probability with kmer length, and sequencing error variances of 3, 9, 15, 21 and 27.

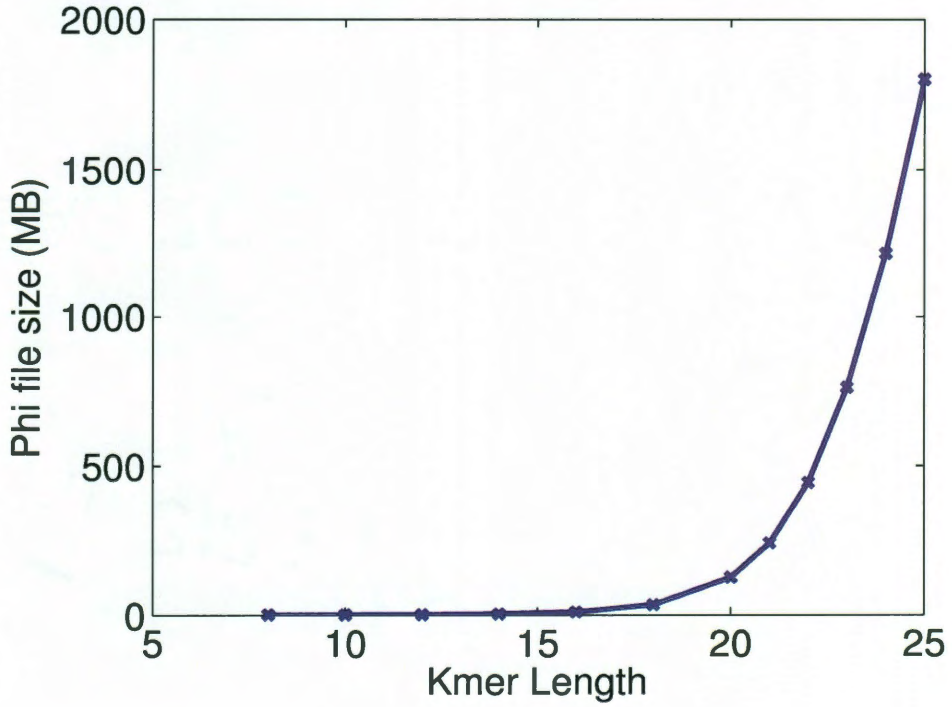


Figure 5.7 : Increasing storage size in MB of Φ with kmer length.

[†] $10\log_{10}(\frac{\text{signal energy}}{\text{noise energy}})$

We see that as kmer length increases, detection performance also improves monotonically, irrespective of the error situation. This is explained by the greater column-wise incoherence that is introduced with increased k, which in turn favors sparsity-based recovery. We know that the RIP (Restricted Isometry Property) is a sufficiency condition on the measurement matrix for most sparsity-based recovery algorithms to guarantee robust error bounds in the presence of noise. For nonnegative matrices in particular it is important to have a sufficient number of zeroes in the matrix such that there is incoherence introduced in the matrix and the RIP is satisfied. In fact, the minimum number of zeroes per row must be lower bounded for Bernoulli matrices to satisfy the RIP [37]. For our purposes, the occurrence of zeroes in the kmer-frequency based Φ stems from having a high enough resolution of kmer bins such that these zeroes are introduced at all. At k = 10, there are 1024 bins almost all of which are nonzero, corroborated by the poor diagonal structure in its Grammian, but at k = 25, even though there are a total of $2^{25} = 33$ million bins possible, only 15 million of them are nonzero for at least one of the 40 bacterial genomes in question. Presumably as the number of bacteria included in the set increase, this number will approach 33 million; however, for any given bacterial genome its size is typically $\sim$ 10 million bases, implying the same number of kmers are available to fill the distribution of 15 million possible kmers generated for our set of 40, leaving a huge number of bins unfilled. Therefore, a rule-of-thumb to choosing a binary kmer length for a given set of bacteria to ensure the fine-grained enough bin resolution that we are after is if,

$$2^k > len_i \quad \forall i = 1...N, \tag{5.2}$$

$$k > \max_i \ \log_2(len_i) \tag{5.3}$$

Here, len_i is the length of the ith bacterial genome in a set of N total. This number

works out to be approximately 23 in the case of E. coli, assuming the double-stranded length of the genome. We see that in practice this is a very loose bound, since starting at kmer length 18 the Figures 5.5 and 5.6 show perfect detection.

It is tempting to choose very high kmer lengths to ensure that our detection ability is good. However, with high kmer lengths come exponentially higher data storage and manipulation needs. Figure 5.7 shows how the memory (in MB) needed to store the Φ's corresponding to different kmer lengths varies. Correspondingly the run times in parsing 100 million sequencing reads against the larger $\Phi's$ also grow rapidly; we saw that generating a 10mer frequency estimate from raw sequencing reads was a matter of seconds, while the 25mer frequency estimate took 2 days on a single 8-core machine using python code without any optimization, while the 10mer frequency estimate took less than a minute on the same machine. A more attractive middle ground is for kmer lengths in between; generating a 20mer frequency estimate took less than 2 hours. It is important to note that the running of the sparsity-based algorithms themselves take comparatively minuscule amounts of time, even for the case of the largest Φ's (on the order of a few seconds to a few minutes).

5.7 Implications of Sparsity-based Kmer Analysis in Sequencing

There are several benefits to using kmer-based estimates combined into a sparse linear model for the analysis of Whole Genome Sequencing data:

1. Increased Analysis speeds

   Key to gaining speed, regardless of kmer length, is the preprocessing of genomes, perhaps at the expense of physical memory, so that there is minimal compu-

tation needed when the data arrives. Our sparsity-based kmer models enable this.

2. Direct application to raw sequencing reads

   Many current sequencing data analysis algorithms require to be applied to assembled contigs in order to interpret differences from reference genomes. Kmer estimate-based analysis method applies directly to raw sequencing reads.

3. Decreased number of sequencing reads

   By using a sparsity-based linear model, we are able to use a much shorter y to decipher x than would otherwise be needed . A decreased number of reads will put less strain on the sequencing process and subsequent computation. (Currently it takes 8-12 hours to sequence 100M 95mer reads, and another 24-36 hours using 20 machines for data analysis with current kmer-profiling algorithms.)

4. Detection over huge target sets

   By using sparsity we are able to ensure that even when the target size grows to be very large – even if on the order of the number of reads available – the number of reads needed for correct target detection will always be far lower than a model that does not assume sparsity.

5. Shorter sequencing read lengths

   We are able to use data from much shorter reads, as long as they are longer than our required kmer length, bucking the current trend toward longer reads (> 90 bases) for more accurate analysis.

6. Innovations in sequencing technologies

There may be enough uniqueness in the single nucleotide skeletons (this is derived when any one nucleotide is set to be a "1" and the others to be "0"'s) of genomes to distinguish between them. This also leads to possible innovations in the sequencing methodology itself; perhaps instead of a complete 4-dye based sequencer only 2 are needed in detection situations – one for the nucleotide of choice, and another for all remaining nucleotides. Furthermore, we may envision a "real-time" sequencing machine where with every cycle we are able to better refine our detection solution. Each cycle adds a nucleotide, allowing us to use a longer kmer length for analysis, and correspondingly allow detection in an equivalence class of species with longer genomes.

7. Positively/negatively selecting for certain bacteria

   It is useful for biologists to identify certain bacteria vs. others in silico, while using the same set of measurements. This could be done by first identifying the full support set of the x vector, and then peeling out the noninteresting bacterial contributions in order to focus on the others.

Chapter 6

Conclusions

In this thesis we have described the application of sparsity-based analysis methods to the problem of bacterial detection in different molecular biology frameworks – microarrays, molecular beacons, sequencing. We see that exploiting the sparsity inherent to a detection problem can confer several different advantages depending on the application at hand.

The bacterial detection problem in almost every situation is a sparse problem, since the number of species in a given sample is always small compared to the large number of sequenced genomes known (several thousand). With sparsity-based tools, in both Compressed Sensing Microarrays and Random Probe Microarrays we saw that we are able to use fewer probes and achieve the same detection results over a large set of bacteria. The probes designed for both these tools work through combinatorial sensing; many overlapping probe patterns collective indicate the presence and quantity of a target. As with all Compressed Sensing measurements, the measurements they take are holistic across the target set, but still differential enough that they can uniquely pinpoint a single target.

The random probe approach to sensing, besides being combinatorial, advocates a paradigm shift in detection altogether: nonspecific sensing. Instead of creating probes that are specific to what we are interested in sensing, we create a probe set that is drawn completely independent of any targets. This approach opens up the possibilities for detection of undiscovered, mutated or newly engineered bacteria, which

would otherwise go unnoticed if specific, tailored probes are used. Given the millions of bacterial species estimated to exist naturally both with harmful and vital purposes, and the potential dangers of ill-purposed synthetic bacterial species being used as bioweapons, it is important for humankind to look to new methods of nonspecific sensing in this domain.

The random probe approach to sensing also holds promise to eliminate PCR. Compared to the conventional unique 16S or 23S ribosomal DNA identifier approach, here not only does each probe oversample the target set by binding to multiple targets, each probe oversamples each genome it binds to by hybridizing in multiple places. It is this multiple probe-binding mechanism throughout a genome that suggests the approach of *whole genome sensing* instead of single gene-based sensing. Random probes bind with genomes on the principle that there are multiple binding spots on each genome which are both unique and shared by various bacteria, and by exploiting them we are able to be flexible in our probe generation method, leading to advantages like nonspecific sensing. The hypothesis that PCR may not be necessary in the case of whole genome sensing is that with enough binding sites on a genome for any single probe type, there is enough cumulative fluorescent intensity that amplification is unnecessary.

Genome-based detection methods are the most definitive way to confirm the presence of a new species. Even today, in clinical environments the standard procedure for the detection of microbial species is culturing. Relying on secondary characteristics of bacteria, even visual inspection, is susceptible to error since it is possible that two different species may still look the same or produce the same antigen. Unfortunately most molecular biology techniques associated with genome-based detection methods are so expensive, long-drawn and fragile that they are many years from being used

outside of laboratory settings. We believe that a more universal approach to sensing with a concept like random probes will ultimately allow for a general species detection device. The ease of a single detection device instead of many, without the need for PCR, could lead to the cost savings that would encourage the wider adoption of genome-based sensing methods.

We have discussed the random probe-based sensing approach in incarnations like microarrays and molecular beacons for microfluidic devices. However, all of these are indirect means of sensing the genome, since our measurements are due to the *piece-wise* complementary hybridization of probes with the DNA to be sensed. A completely different approach to genomic sensing may be given by the adoption of sequencing technology. If genome-based methods are the gold standard for species detection, sequencing is the gold standard among genome-based methods, since it allows us to literally *read* the genome, base-by-base.

There have been monumental and rapid developments in diverse sequencing approaches over the last 6-8 years, collectively labeled as Next Generation Sequencing methods. While currently exorbitant, the competition and diversity of commercial sequencing approaches ensure that its price will be driven down and its accuracy will go up. These changes will make a difference in deciding whether and how soon sequencing-based technologies will be available for genomic detection purposes – whether bacterial species or humans.

This thesis discusses the use of sparsity-based methods in the bacterial species detection from Whole Genome Sequencing (WGS) data, where PCR amplification is typically not necessary since the entire genome is sequenced instead of a few genes. WGS is frequently used to identify species in diverse microbial communities such as an environmental sample or in the human gastrointestinal tract. However, this

approach currently faces computational challenges due to the long analysis times that many unique-profiling methods need for high accuracy rates. Instead, we make the observation that the detection problem here is in fact a sparse detection problem as we have solved before, and apply sparsity-based analysis tools in a linear model. This approach allows for maximum time spent in preprocessing, and consequently shorter analysis times using the actual data. By recognizing the sparsity in our model, we are able to exploit flexibilities that require fewer sequencing reads for the same accuracy. If minimizing the number of reads is not a goal (as in many sequencing applications) then the use of additional reads contributes to further improvements in recovery accuracies.

One major shortcoming of sequencing technologies is the data deluge that they produce – terabytes of sequencing data quickly accumulate and are computationally combed through to establish biological facts. It is therefore imperative to simplify this process as much as possible; either by limiting the amount of data produced for each specific application, or limiting the amount of data needed for the interpretations sought computationally. The assumption of a sparsity-centric data model and the associated analysis tools that leverage it help on both these counts.

Over the last five years since the emergence of Compressed Sensing theory and sparsity-based methods, there has been a surge of applications where sparsity is used as analysis tool but very few newly engineered devices that exploit it. This is not in fact pathological to the field of CS. Generally, signal processing and data analysis methods are more frequently used to better interpret data than as platforms from which to build more intelligent data-gathering systems. In this thesis we have tried to do precisely this – to use sparsity to create new real world tools with previously unseen advantages. We raise the need for more instances of this type of reverse

influence, where analysis methods inform better instrument development in the first place. This is not practical in many situations, but in special cases – like sparsity and Compressed Sensing – there are a multitude of practical, translatable advantages that will enable us to engineer better technologies for tomorrow.

Bibliography

[1] W. Weisburg, S. Barns, D. Pelletier, and D. Lane, "16S ribosomal DNA amplification for phylogenetic study," *Journal of Bacteriology*, vol. 173, pp. 697–703, Jan. 1991.

[2] E. J. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Info. Theory*, vol. 52, pp. 5406–5425, Dec. 2006.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, pp. 1289–1306, Sept. 2006.

[4] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.

[5] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, March 2001.

[6] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, 2008.

[7] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Info. Theory*, vol. 53, Dec. 2007.

[8] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26,

pp. 301–321, 2008.

[9] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Compressed sensing reconstruction via belief propagation," Rice University Tech Report ECE-06-01, Oct. 2006.

[10] R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Info. Theory*, vol. 56, April 2010.

[11] A. Katsnelson, "Epigenome effort makes its mark," *Nature News*, Oct. 2010.

[12] D. Shalon, S. J. Smith, and P. O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Research*, vol. 6, pp. 639–645, 1996.

[13] "www.molecular-beacons.org."

[14] M. L. Metzker, "Sequencing technologies – the next generation," *Nature Reviews Genetics*, vol. 11, Jan. 2010.

[15] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, 1981.

[16] J. SantaLucia Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.," *Proc. Natl. Acad. Sci.*, pp. 1460–1465, 1998.

[17] J. SantaLucia Jr. and D. Hicks, "The thermodynamics of DNA structural motifs," *Annual Rev. Biophys. Biomol. Struct*, vol. 33, pp. 415–440, 2004.

[18] Y. Chen, C. Chou, X. Lu, E. Slate, K. Peck, W. Xu, E. Voit, and J. Almeida, "A multivariate prediction model for microarray cross-hybridization.," *BMC Bioinformatics*, vol. 7, no. 1, pp. 101–112, 2006.

[19] I. Hofacker, "Vienna RNA secondary structure server.," *Nucleic Acids Research*, vol. 31, pp. 3429–3431, 2003.

[20] W. Dai, O. Milenkovic, M. A. Sheikh, and R. G. Baraniuk, "Probe design for compressive sensing DNA microarrays," in *IEEE Int. Conference on Bioinformatics and Biomedicine*, 2008.

[21] M. A. Sheikh, S. Sarvotham, O. Milenkovic, and R. G. Baraniuk, "DNA array decoding from nonlinear measurements by belief propagation," in *IEEE SSP Workshop*, (Madison, WI), Aug. 2007.

[22] M. A. Sheikh, O. Milenkovic, and R. G. Baraniuk, "Designing compressive sensing DNA microarrays," in *Second International Workshop on Computational Advances in Multi-Sensor Adaptive Processing.*, 2007.

[23] W. Dai, M. A. Sheikh, O. Milenkovic, and R. G. Baraniuk, "Compressive sensing DNA microarrays," *EURASIP Journal on Bioinformatics and Systems Biology*, 2009.

[24] A. Schliep, D. Torney, and S. Rahmann, "Group testing with DNA chips: Generating designs and decoding experiments," in *Proc. of Computational Systems Bioinformatics Conf.*, 2003.

[25] D. Z. Du and F. K. Hwang, *Combinatorial group testing and its applications.* World Scientific Publishing Co., 2000.

[26] "Clusters of orthologous groups — NCBI/NIH," Available at http://www.ncbi.nlm.nih.gov/COG/.

[27] B. Snel, P. Bork, and M. A. Huynen, "The identification of functional modules from the genomic association of genes.," *PNAS*, vol. 99, no. 9, pp. 5890–5895, 2002.

[28] D. Hekstra, D. Taussig, A. Magnasco, and M. Naef, "Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays.," *Nucleic Acids Research*, vol. 31, pp. 1962–1968, 2003.

[29] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, pp. 1207–1223, Aug. 2006.

[30] M. A. Herman and T. Strohmer, "General deviants: An analysis of perturbations in compressed sensing," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Compressive Sensing*, vol. 4, pp. 342–349, April 2010.

[31] D. B. Polk and R. M. Peek, "Helicaobacter pylori: gastric cancer and beyond," *Nature Review Cancer*, vol. 10, 2010.

[32] S. Chatterji, I. Yamazaki, Z. Bai, and J. A. Eisen, "Compostbin: a DNA composition-based algorithm for binning environmental shotgun reads," in *Proceedings of the 12th annual international conference on Research in computational molecular biology*, RECOMB'08, (Berlin, Heidelberg), pp. 17–28, Springer-Verlag, 2008.

[33] H. Zheng and H. Wu, "A novel LDA and PCA-based hierarchical scheme for metagenomic fragment binning," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2009*, pp. 53 –59, 2009.

[34] M. Rosenblatt, *Stationary Sequences and Random Fields*, pp. 196–199. BirkHaüser, 1985.

[35] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, 2009.

[36] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[37] V. Chandar, "A negative result concerning explicit matrices with the restricted isometry property." Preprint, 2008.